Measuring Cannibalization in Distribution Networks: an Approach to Optimize Store Locations

By

Sergio C. Sapaj

Master of Science in Marketing, Johns Hopkins University (2011)
Bachelor of Science in Industrial Engineering, Universidad Técnica Federico Santa María (2004)

SUBMITTED TO THE SYSTEM DESIGN AND MANAGEMENT PROGRAM IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2013

Signature of Author ____________________
System Design and Management Program
January 10, 2013

Certified by ____________________
Professor Roy E. Welsch
Thesis Supervisor
Professor of Management Science, Statistics, and Engineering Systems
Sloan School of Management and Engineering Systems Division

Accepted by ____________________
Patrick Hale
Director
System Design and Management Program

This page has been intentionally left blank

Measuring Cannibalization in Distribution Networks: An Approach to Optimize Store Locations

By

Sergio C. Sapaj

Submitted to the System Design and Management Program on January 15, 2013 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Engineering and Management

Abstract

A methodology was proposed to measure sales cannibalization using historic data containing variations in store densities. Sales cannibalization was defined as a decrease in the sales of one or several existing stores as a result of nearby store openings. The proposed methodology, combining regressions with geographic and other forms of cluster analysis, allowed measurement of the cannibalization, while controlling for other relevant sales drivers such as consumers' and geographic areas' characteristics, seasonality and the nature of the demand of the product category (impulsive and non-impulsive purchases). The analysis found evidence of sales cannibalization in the store network studied and showed that its severity varies according to consumers' and geographic areas' characteristics and product category. This last finding was particularly relevant, as cannibalization was consistently more severe for non-impulsive products.

Based on the cannibalization measurements, sales functions were created and then optimized to find the number of stores maximizing total sales. This number represents saturation, meaning a point beyond which any new store opening in the area just redistributes sales. The number of stores maximizing sales, however, may not be the goal, particularly when fixed costs associated with operating stores are important and when attempting to maintain attractive businesses for store owners, which is relevant in franchised settings.

Thesis Supervisor: Professor Roy E. Welsch
Title: Professor of Management Science, Statistics, and Engineering Systems, Sloan School of Management and Engineering Systems Division, Massachusetts Institute of Technology

This page has been intentionally left blank

Acknowledgments

I would like to thank Professor Welsch for his guidance, help and support throughout the process of developing this thesis. His classes were also particularly helpful to develop the analyses and methodologies required to achieve this work.

I would also like to thank the System Design and Management (SDM) staff; they have been very helpful during the program. I must thank Pat Hale for his constant support not only to develop this thesis, but also in many other requests and endeavors.

My classmates and friends at MIT have made my experience enjoyable, memorable and rewarding. I learned a great deal from you guys and I am definitely going to miss you.

Finally, I would like to thank my wife, Alejandra, for her love, support and advice. Without you none of this would have been possible.

Table of Contents

Figure Index

Table Index

Introduction

Designing and optimizing distribution channels is perhaps one of the most crucial tasks that companies with extended distribution networks need to perform to ensure their profitability. Good location decisions will result in higher revenues and profits, while bad ones could become financial liabilities (Kimes et al., 1990). Opening stores involves capital expenditures and, therefore, if locations are not properly chosen, there will be a financial burden associated with having to relocate them (Achabal et al. 1982).

As distribution networks grow, the effect that new store openings will have on the established network becomes a critical issue because locations that used to be profitable may no longer be due to redistribution effects (Garee et al., 1998). In the "real world," however, decisions concerning opening new stores or closing or strengthening current ones are usually made disregarding this issue, as, at most, "myopic" models, i.e., models that optimize the location selection for the next store to open, are used, without considering the effect that newly added stores will have on the rest of the network (see Craig et al. for a complete summary of different models to define store locations).

When location models are used, there is usually a trade-off among companies between the use of more complex methods and the frequency and investment associated with the openings. Companies with large quantities of rather small stores (e.g., mom and pop type stores), rely mostly on employees' judgment to make location decisions, as opposed to, for example, department stores or shopping malls, on the other extreme of the spectrum, which often use some sort of systematized tool. The latter ones undertake fewer location decisions per unit of time, involving large and expensive development projects, which frequently start with some form of optimization analysis to find the "right" location. Companies with large quantities of smaller stores, on the other hand, open and close them more frequently, involving smaller investments and quicker projects, where decisions are habitually made by mid-low level employees in the "search for growth," usually without any formal analysis. This is the case also of companies not operating their own stores, but working with existing third party ones. In these cases, it is frequently a salesman who enrolls new stores for his company, offering usually in exchange additional infrastructure for the store owner (e.g., soft drink companies providing existing stores with additional cooling infrastructure). In both cases, however, the effect that new stores will have on existing ones is not properly addressed as even the more complex models are based on customers' surveys and secondary data (Durvasula et al. 1992), not considering historic data systematically to learn from past openings and their effect on the existing network. There are some models (see Ghosh et al., 1991 and Kaufman et al., 1990) that have tried to expand single location methods to optimize sales from a more holistic perspective, considering also existing stores in the case of franchised settings, but they are still based on consumer surveys and do not use relevant historic data. Survey-based methods are more expensive to implement, have to deal with systematic and non-systematic errors and biases arising from sampling mechanisms and questionnaires and disregard evidence of the sought behavior in history. While in a survey-based method other factors, like increases in the areas' attractiveness as the result of more stores for example, have to explicitly be considered and incorporated into the analysis through a series of assumptions (Ghosh et al., 1991), a history-based approach deals with all the effects implicitly by considering the results observed in past data, which already incorporate relevant dynamics.

In a distribution network, the revenue produced by any new store comes from three streams:

- New sales generated by providing consumers with additional opportunities to access the company's products.
- Sales captured from competitors, operating either in nearby locations, in the case when a new store is being opened (e.g., McDonald's capturing sales from a nearby Burger King), or within the store, when the company is enrolling existing ones (e.g., Coca-Cola capturing sales from Pepsi when a store that originally was selling only Pepsi products starts to sell products from both companies).
- Re-allocation of company's sales, from nearby stores to the new ones.

When opening a new store, companies should seek locations where the third revenue stream, usually known as cannibalization, is low enough to maintain both distribution efficiencies and a business sufficiently attractive for the current stores in the area (e.g., if Coca-Cola's sales in a given store decrease after a new nearby store is enrolled, it may no longer be attractive for the store owner to carry Coca-Cola's products, and he may decide to allocate that space for a new product category). The latter effect is particularly relevant for companies operating with exclusive distributors, usually requiring high investments.

Acknowledging these difficulties, this thesis proposes a mechanism to guide new store openings based on cannibalization measurements obtained from historic data, which is conceptually more precise and easier and cheaper to obtain than survey data. The results of this analysis, as will be shown, clearly demarcate opportunities for growing in the territory selected, indicating also areas where store saturation has reached dangerous thresholds, requiring company's actions such as store reallocations or increases in marketing expenditures in the area (that could lead to demand generation hindering the negative effects of cannibalization). The analysis is based on the information about store openings and closings contained in historic data, and how changes in stores' densities affected sales of existing stores.

Chapter 1: Methodology

The aim of this thesis is to propose a mechanism to guide store location decisions based on evidence of sales cannibalization in historic data. The first step then is to adopt a cannibalization definition to be used in the rest of the thesis. Accordingly, cannibalization is defined for the purposes of this thesis as "the decrease in the sales of a store as a result of the addition of one or more stores to the distribution network near it." This definition is consistent with what is found in the literature (Garee et al., 1998 and Ghosh et al., 1991 use similar definitions).

The process started by geocoding stores' addresses (i.e., converting stores' addresses into coordinates). Cannibalization was expected to differ among product categories, with products whose purchase process is mostly driven by impulse being less prone to cannibalization than those with more planned purchases. Accordingly, geocoded stores were placed in two different layers: one layer associated with the impulsive product and another layer associated with the more planned purchase product. Stores selling both categories were placed in both layers and treated as independent stores sharing an address. This separation was kept throughout the process, and independent analyses for both product categories were performed. Only in the last stage of the process were both layers superposed with the purpose of finding common areas of opportunities.

Once addresses were geocoded and both layers were created, the next step was identifying stores that were near each other. This was done by grouping stores in "store neighborhoods" or "bubbles," formed by drawing circles with different radiuses around them, and then merging overlapping circles into one store geocluster (name subsequently interchangeable used to refer to store neighborhoods or bubbles). By creating geoclusters using different radiuses, layers with alternative definitions of what was "near" were considered in the analysis.

Cannibalization was expected to depend not only on the product category, but also on the demand potential of the areas covered by the store geoclusters. Accordingly, store geoclusters were characterized in terms of the demand profile they served, which was done through a clustering process of the city's census tracts: census tracts were clustered in 15 groups according to variables such as inhabitants' income level, population density and pedestrian traffic. Store geoclusters were then intersected with census tracts, inheriting their demand clusters. This process allowed classifying store geoclusters in groups with homogeneous demand potential (demand clusters). Subsequent cannibalization analyses were performed within demand clusters, as the effect of cannibalization was expected to be different. It is pertinent to note that the variables chosen for the demand clustering process were selected according to their expected influence on the demand for the product categories; accordingly, once the clustering was completed, an analysis of the resulting clusters was conducted to test the capability of the clustering process to add demand homogeneity. The results of the analysis confirmed that clusters indeed reduced demand heterogeneity substantially.

With store geoclusters grouped in homogenous demand clusters, the next step was to calibrate functions relating changes in the number of stores with variations in stores' sales. These functions contain the definition of cannibalization. To build these functions, the average number of stores and the average sales of the stores within each geocluster were calculated; the functions then linked deviations from the mean in the number of stores with deviations from the mean in the average sales of the stores. The hypothesis was that as more stores are added to geoclusters (a positive deviation from the mean of the number of stores), due to cannibalization, the average sales of each store within geoclusters should decrease (a negative deviation from the mean in the sales of the stores). This analysis was done using several functional forms (linear, exponential and a differences model), within each demand cluster, radius layer and product category, adjusting functions through least squares and then testing the robustness of the estimated parameters using 500 sample bootstraps. In every case, coefficients controlling for seasonal effects were included.

Once the functions linking changes in average sales with variations in stores' density were calibrated, using algebraic manipulations, sales functions for the supplier company were built for the store geoclusters, in which total sales were expressed as a function, among other variables, of the number of stores. These functions were then maximized with respect to the number of stores, to find what is referred to in the last chapter as the saturation number of stores, meaning the number of stores that maximize the company's sales in each geocluster (implying that every new store added to the geocluster beyond this point only redistributes sales). By comparing the currently active with the saturation number of stores, gaps to saturation were calculated. These gaps were then recalculated using the extreme values for the parameters of the functions found through the bootstrap to obtain a sense of the risk (i.e., the probability of going beyond the saturation point when using the set of coefficients derived directly from least squares) associated with pursuing strategies aimed at closing the gaps. This gap analysis offers guidance to management regarding areas of the city that can still support more stores, compared to others already saturated, in which new store openings will most likely only redistribute sales instead of supplying new sales.

Chapter 2: Geoclustering

Attending to the definition of cannibalization that is being used for this thesis (i.e., "a decrease in the sales of a store as a result of the addition of one or more stores to the distribution network near it"), it was necessary to identify stores that are "near" each other.

To identify nearby stores, a form of geographic clustering was conducted in which stores near each other were grouped in clusters. The process began by drawing circumferences with different radiuses[1] around each of the stores:

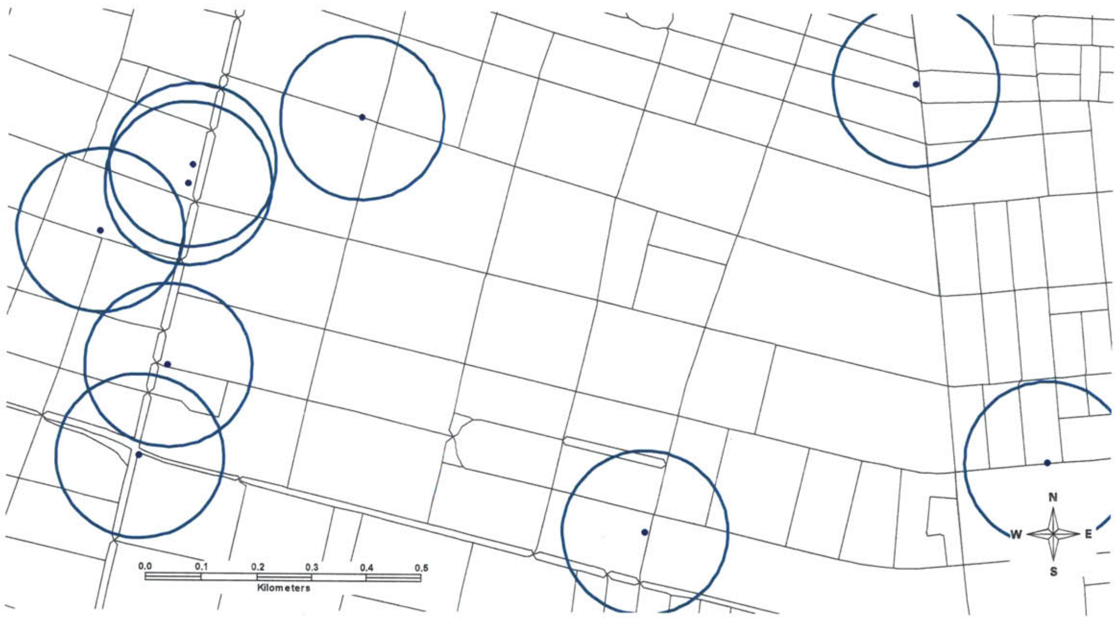

Figure 1: First Step, Forming Circumferences, Geoclustering Process

Then, touching circumferences where merged, to form "store geoclusters" (circumferences without intersection with others became store geoclusters with only one store in them):

[1] Later in the chapter more information is provided regarding this point.

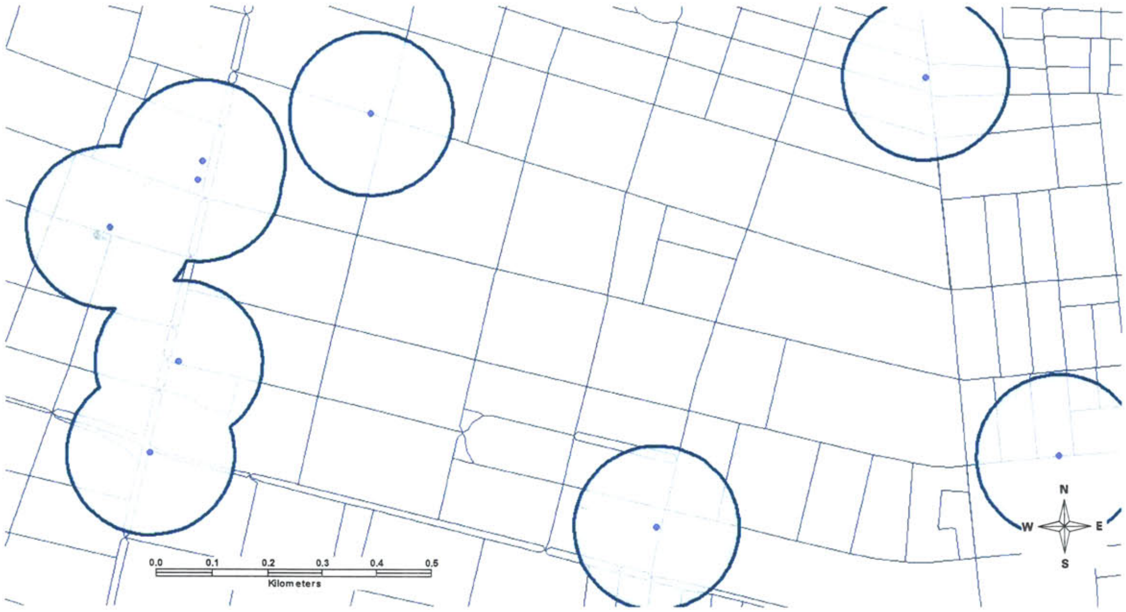

Figure 2: Second Step, Merging Touching Circumferences, Geoclustering Process

Given the definition of cannibalization used, the main hypothesis was that variations in the density of active stores within geographic clusters will cause variations, in the opposite direction, over the average size of the stores within them. Accordingly, it was required to consider for the geographic cluster analysis the entire universe of stores with at least one transaction in the analysis period (from September 2009 to December 2010 for lottery games, and from January 2010 to January 2011 for instant games). In this regard, one layer (per radius) of geographic clustering was built for the entire period for each product category; what varied, as stores activated and deactivated, was the clusters' store density.

This process was repeated with different radiuses. The following table summarizes the results:

Table 1: Summary of Geoclusters

Radius [m]	N° of Clusters
50	1,025
100	857
150	694
200	542
300	294
400	112

A secondary hypothesis emerges from the use of different radiuses: as the store geoclusters' radiuses vary, the expected magnitude of the cannibalization phenomena should also vary. In this regard, bigger radiuses imply geoclusters covering larger areas, with more demand potential (assuming a homogeneous density for demand potential measured, for example, in \$/month*m^2, bigger areas, as a result of bigger radiuses, would deliver more demand to serve, in \$/month), which should allow them to support more stores with lower cannibalization.

The following map shows an example of the results obtained for the geoclustering process with 200 m. The color of the geoclusters represents the number of stores included in it (lighter blues are associated with higher numbers of stores):

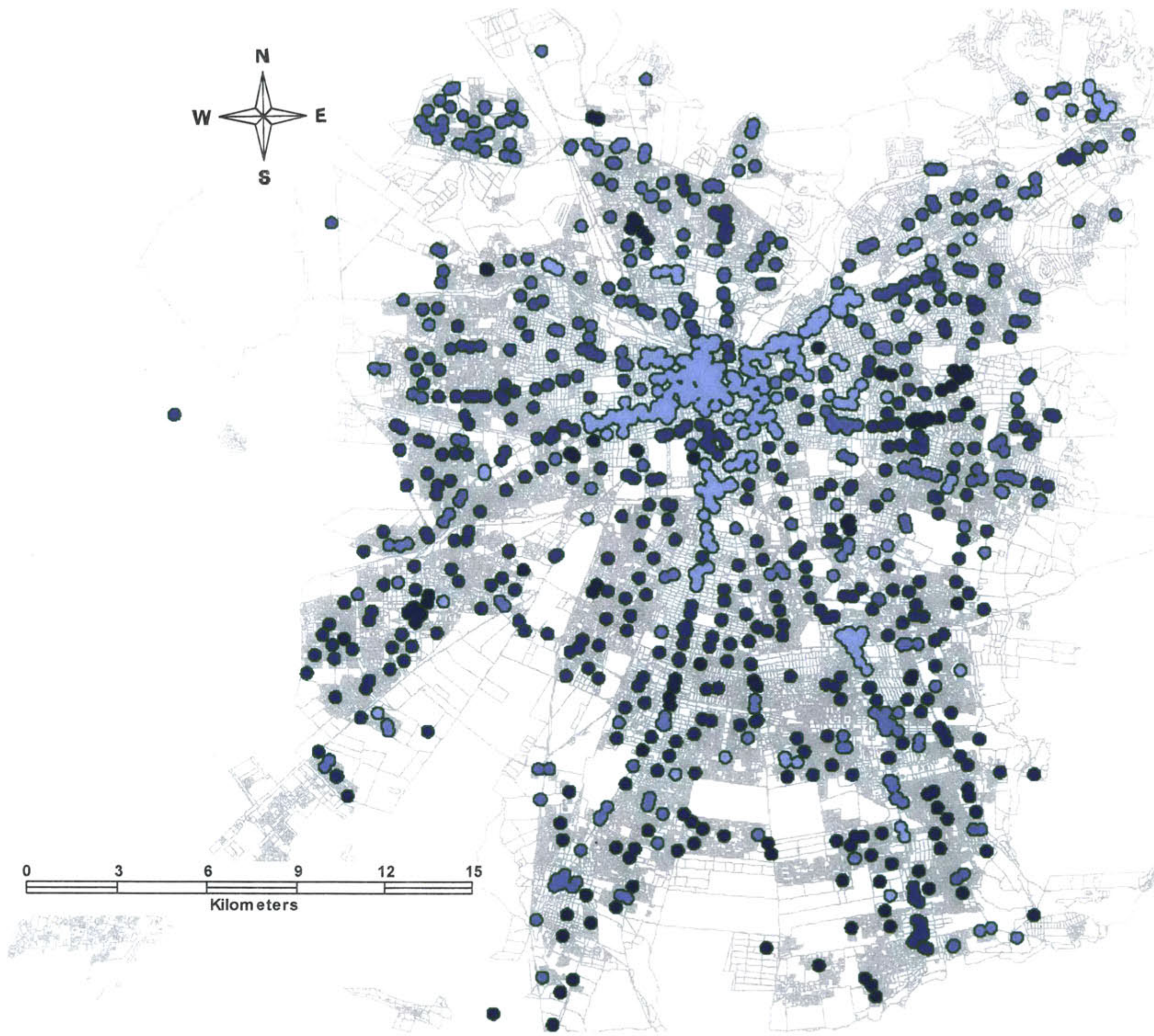

Figure 3: Results of the Geoclustering Process, 200 m

As seen in the map, there are several store geoclusters with only one store (the deepest blue circles). The following table shows the distribution of the number of stores per geocluster for the case of different radiuses:

Figure 4: Summary of the Number of Stores per Geocluster

	Number of Cases					
Number of Stores	50 m	100 m	150 m	200 m	300 m	400 m
1	848	654	484	322	124	46
2	124	127	121	120	67	17
3	32	32	36	43	28	9
4	10	26	19	18	25	9
5	3	7	10	10	9	5
6	2	2	5	9	8	4
7	2	2	4	4	9	4
8	2	3	1	4	4	2
9	0	0	3	0	4	1
10	1	0	3	5	3	0
11	1	2	3	0	2	2
12	0	1	0	0	1	1
13	0	0	0	2	2	1
14	0	0	0	0	2	2
15	0	0	2	0	0	1
16	0	0	0	0	0	2
17	0	0	0	0	0	1
18	0	0	1	2	0	0
19	0	0	0	1	1	0
20+	0	1	2	2	5	5
% Uni-store clusters	83%	76%	70%	59%	42%	41%
Total	1025	857	694	542	294	112

As expected, as the radius increases, the average number of stores per geocluster also increases. In this regard, small radiuses, like 50 m for example, have almost all the stores assigned to uni-store (i.e., geoclusters with only one store) geoclusters, implying that they would provide fewer observations with variations in the number of stores per cluster to measure its effect on cannibalization. On the other hand, bigger radiuses, like 400 m for example, would provide more observations with variations in the number of stores, but their geoclusters would be more heterogeneous, as they cover several potentially different census tracts, with different demand characteristics. To balance these two effects, the geoclustering layers used in subsequent analysis were 150 m, 200 m and 300 m. Although these layers were still heavily dominated by uni-store clusters, these observations were still useful as they helped to estimate the effect of seasonality in the analysis.

Chapter 3: Characterizing Demand Using Census Tract Cluster Analysis

In order to obtain meaningful results from the analysis, it was necessary to control for the areas' demand potential served by each geocluster. In this regard, cannibalization should be less important, all other things being equal, in areas where there is a stronger demand for the products or services offered by the company, because there is higher revenue potential to support stores.

To control for this effect, store geoclusters were grouped according to the demand potential of the geographic area they cover. To create these groups of homogenous geoclusters, the census tracts they touch were characterized according to relevant demand traits, and then accordingly clustered. Geoclusters were assigned later to clusters based on their intersection with census tracts. In the event that a geocluster intersected with census tracts belonging to different clusters, the geocluster was assigned to the cluster with which it had the greatest area of intersection.

All the analyses were done using data from the distribution network of the Chilean National Lottery Company (further referred as the "company"). Company' sales were grouped into two product categories: instant games and lottery games.

Variable Selection

Demand for products or services can be characterized by three groups of variables:

- Demographic variables: when assessing demand characteristics, consumers' income level, age profile and gender are usually the most important variables within this group.
- Situational variables: regardless of the income level or age profile, consumers usually show different consumption habits depending on the situation in which they are: at home, at school, on the way to work, etc.
- Psychographic variables: these are variables related to consumers´ characteristics that are relevant for the products' demand but are not covered with demographics. They are related to consumers' lifestyle, tastes and preferences and habits.

The combination of these three variables groups allows a more profound understanding of the demand potential faced by each store. For this thesis, however, only variables regarding the first two groups were available:

- Demographic information, available at the level of blocks, from the 2002 Chilean National Population Census.
- Situational variables, captured indirectly through points of interest (POI) in the cartography, such as schools, hospitals, ATMs, banks, subway stations, bus stops, shopping malls, and other relevant point of interest. In this regard, what were measured were the drivers that make people move within the cities, not the actual flow of people (which is expensive and hard to measure).

The third variable group, psychographic information, was not readily available at the moment this thesis was being developed.[2] Nevertheless, the assumption is that part of its effects are covered by the other two variable groups, and people with different psychographic characteristics are equally likely to purchase in any store within a given cluster assembled considering only demographic and situational information.[3]

Census Tracts as Unit of Analysis

Census information was available at the level of blocks. It was aggregated at the census tract level before doing the cluster analysis. This aggregation was made to have a more homogeneous unit of analysis, avoiding the variability that working at the block level would have brought. There are 1,138 Census tracts in Santiago, having an average surface of 726,777 m^2 and containing, in average, 41 blocks. The following map shows an example for two counties in Santiago (Providencia and Ñuñoa). Red lines show census tracts limits, gray lines show blocks and black lines the limits of the counties.

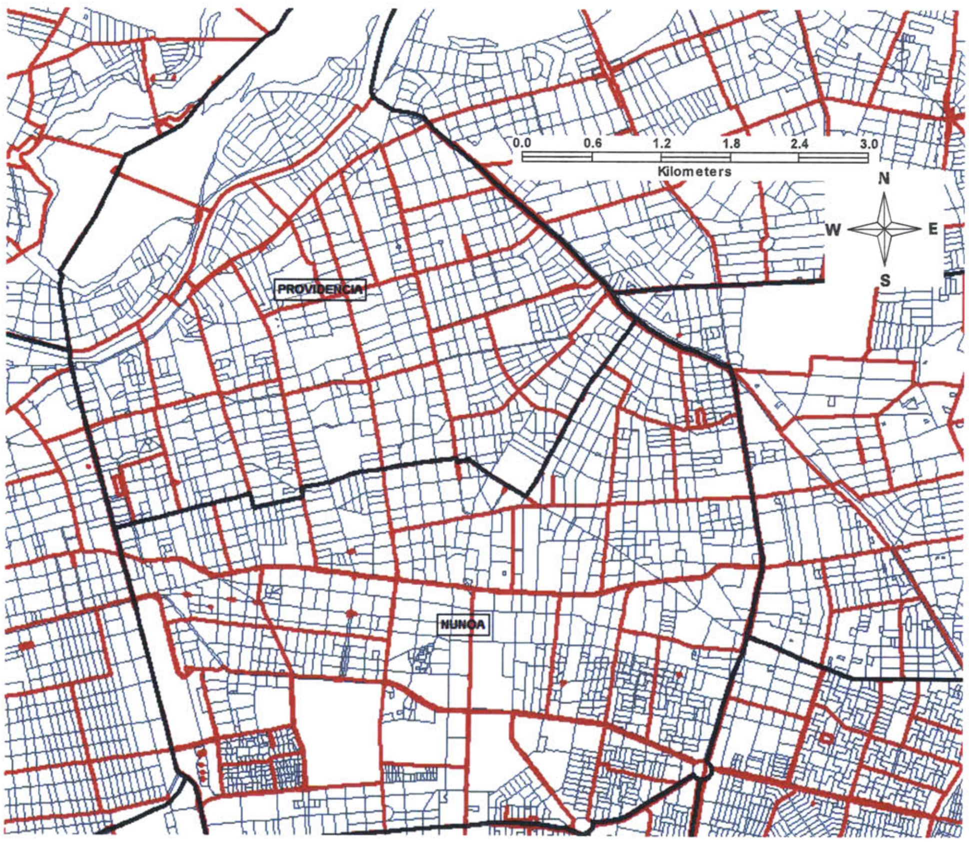

Figure 5: Census Tract View

[2] To get information regarding this group of variables implies efforts beyond the scope of this thesis.

[3] This assumption is going to be tested later on this thesis through an analysis of the errors produced by the regression used to build the cannibalization models. If the lack of this variable group biases the results, patterns should be observed in the models' residuals.

For each of the census tracts, and based on, as was mentioned, census information available at the block level, the following variables were calculated:

- Number of families per socio-economic group[4] as a proxy of the income level: ABC1 (the highest group), C2, C3, D and E (the lowest).
- Total number of people living in the census tract.
- Census tract's population density, in people/m^2.

Census tracts were also analyzed using cartographic and POI information. In this regard, variables measuring the number of each of the following entities were calculated: drugstores, banks, ATMs, shopping malls, supermarkets, subway stations, bus stops, gas stations, important avenues, health centers, hospitals, police stations, schools and universities. Bus stops were further broken down into local, long-run and transfer stops.

Variable Reduction

The next step was to create groups to reduce the number of variables for the census tract cluster analysis. Accordingly, the following variables and indexes were created (some of them weighting raw variables):

- Census tract main socioeconomic group:
    - High: census tracts mainly with ABC1 households.
    - Medium High: census tracts with mainly C2 households.
    - Medium: census tracts with mainly C3 households.
    - Medium low: census tracts with mainly D households.
    - Low: census tracts with mainly D and E households.
- Commerce index (weighted score): banks + drugstores + ATMs + 2 x supermarkets + 4 x malls.
- Transportation index (weighted score): local bus stops + 3 x transfer bus stops + 2 x long run stops + 4 x subway stations + avenues + gas stations.
- Service index (weighted score): 2 x police stations + 2 x health centers + universities + schools + 4 x hospitals.

Weights (e.g., 2, 3 and 4) were used to calculate indexes to capture their different capabilities to attract pedestrian traffic. Weights' values are based on experience of similar projects done in the past by the author of this thesis, although the cluster analysis was not particularly sensitive to their values.

The following statistics were obtained for each variable that was included in the cluster analysis:

[4] According to the criteria used by the Chilean Association of Market Research Companies (www.aimchile.cl).

- Census tract main socioeconomic group:

Table 2: Census Tracts Socioeconomic Group's Frequency

Group	Frequency	Percentage
Low	385	33.8
Medium low	498	43.8
Medium	67	5.9
Medium High	115	10.1
High	72	6.3
Total	**1,138**	**100**

- Commerce index:

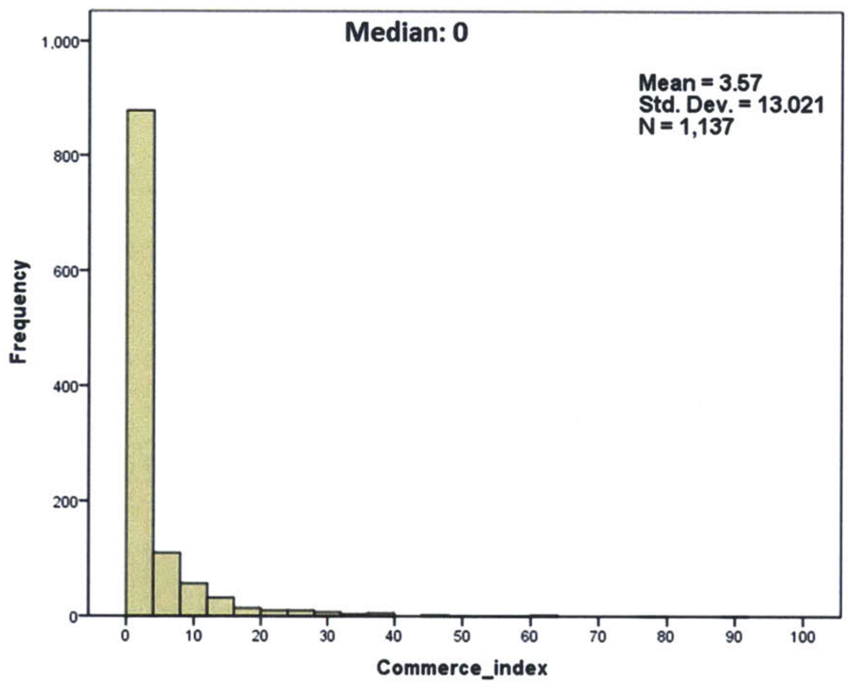

Figure 6: Commerce Index Distribution

- Transportation index:

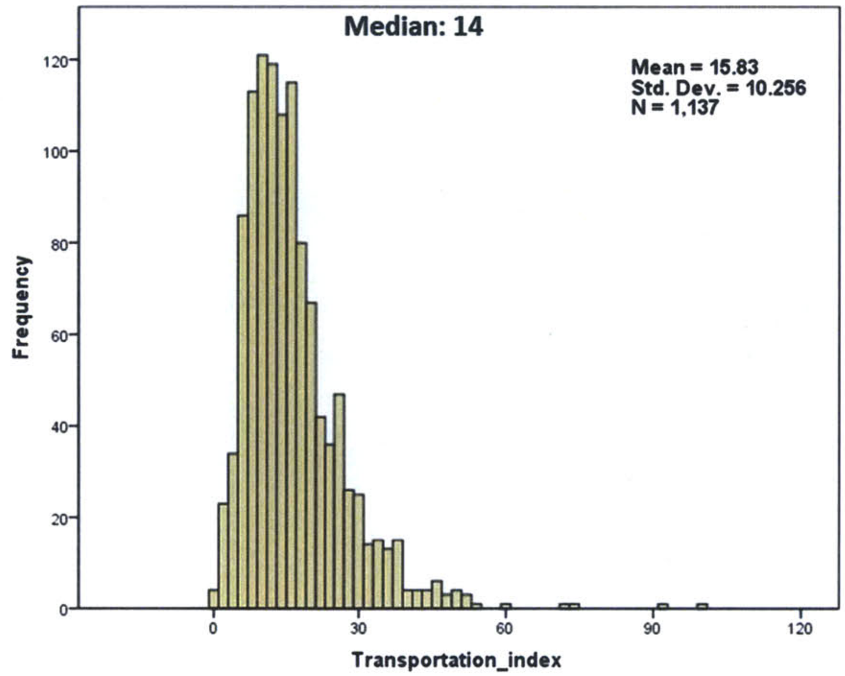

Figure 7: Transportation Index Distribution

- Service index:

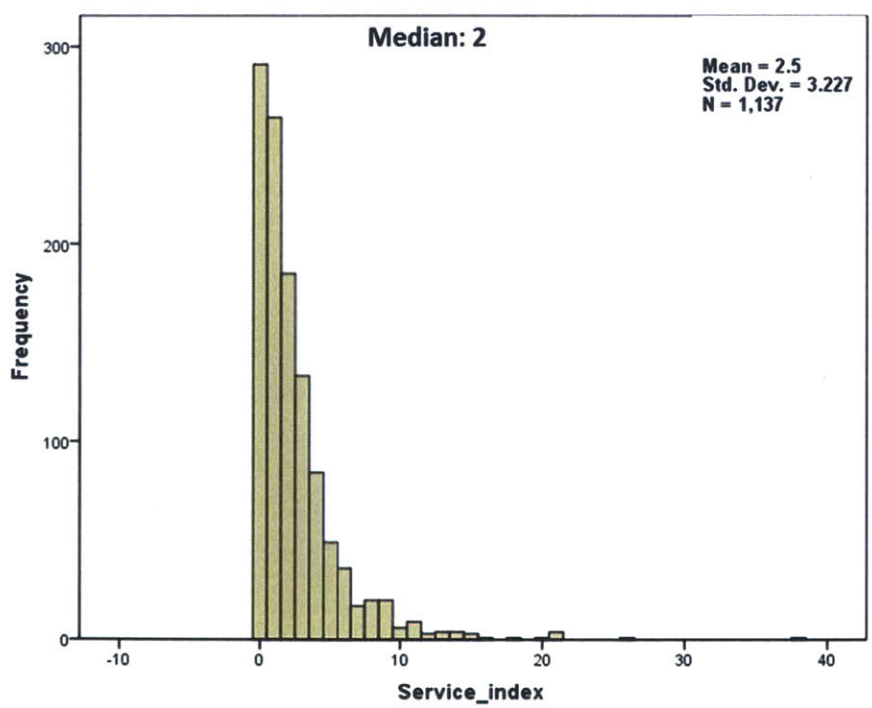

Figure 8: Service Index Distribution

- Population density, in people/m^2:

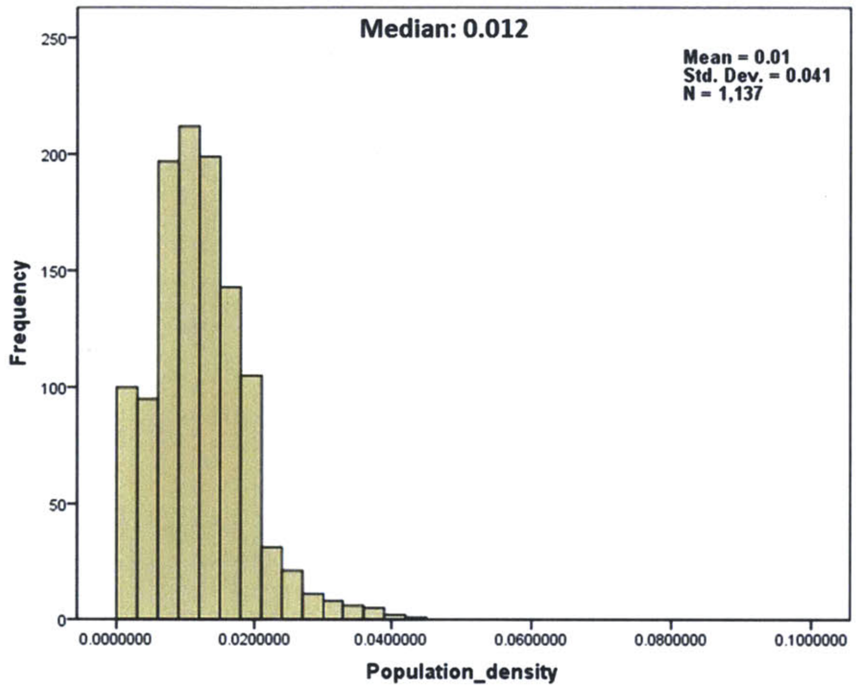

Figure 9: Population Density Distribution

- Total Number of People:

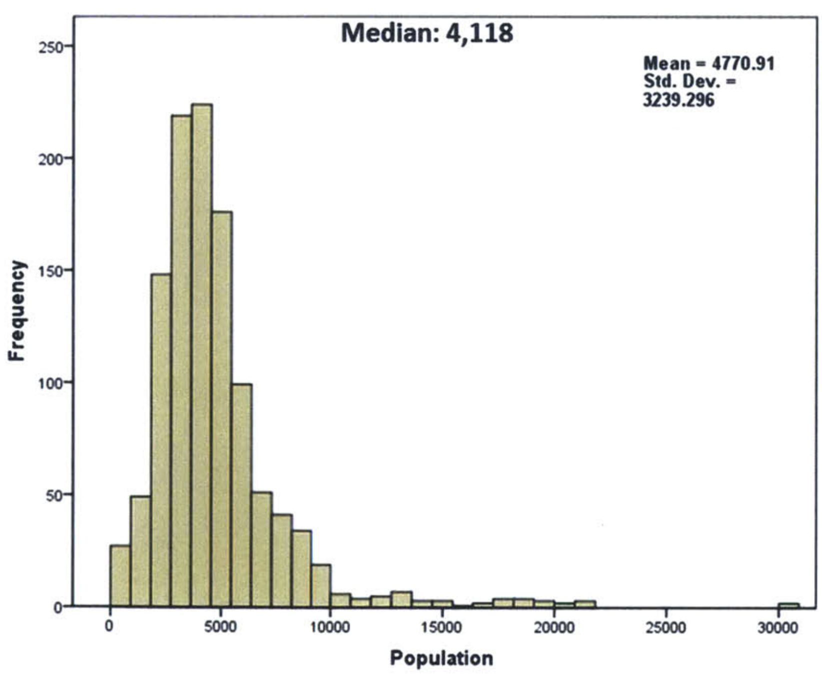

Figure 10: Population Distribution

Different variables had different scales and distributions. To avoid any scaling issues during the clustering procedure, all variables were normalized[5] before using them.

Cluster Analysis

As mentioned before, the objective of grouping census tracts into clusters was to have more homogeneous units of analysis regarding the phenomenon being studied (sales cannibalization). Accordingly, variables were selected under the hypothesis that census tracts sharing both demographic and situational characteristics would have similar demand potential for the products sold by the company (both instant games and lottery games). At the end of this section a procedure is proposed to test this hypothesis by measuring the ability of clusters to generate sales homogeneity.

The cluster analysis produced 15 clusters that were not the direct result of a single run with a given cluster algorithm; instead, a recursive procedure was followed observing the trade-off between the number of tracts in each cluster and homogeneity.[6] An initial clustering run was executed, and then the results were analyzed both in terms of homogeneity and number of census tracts per cluster:

- For clusters where heterogeneity was still high, new clusterings were run considering only their census tracts. If two or more of those highly heterogeneous clusters were "close" enough to each other, the new clustering run was executed considering their census tracts altogether.
- Clusters that were too small (in terms of the number of census tracts they contained) were mixed with their "closest" neighbor, and the resulting new cluster was analyzed in terms of homogeneity. If its homogeneity was still within acceptable ranges, the cluster was considered final; if not, a subsequent cluster run was executed within the consolidated cluster.

This procedure was recursively repeated several times until 15 clusters were obtained, representing a balance between a minimum number of census tracts per cluster to ensure reliable posterior analysis, and homogeneity to make analysis within clusters meaningful. The following table summarizes clusters' centroids (the variable "Census tract main socioeconomic group" is not shown; instead, the percentages of households for each raw socioeconomic group are displayed). Each of them was labeled with a description based on its main characteristics:

[5] Normalizing variable X through

[6] The following section will explain the index used to measure homogeneity.

Table 3: Summary of Clusters' Centroids

Cluster Description	People [Person]	Density [Person/10,000 M^2]	Commerce Index	Transportation Index	Service Index	ABC1 Households [%]	C2 Households [%]	C3 Households [%]	D Households [%]	E Households [%]
1. High socioeconomic level, high concentration of commerce, transportation and service points, low population density	5,943	59	9	20	3	55%	32%	9%	3%	1%
2. Medium-high socioeconomic level, with concentration of commerce, transportation and service points	4,617	86	10	22	5	28%	41%	20%	10%	1%
3. Medium socioeconomic level, high concentration of transportation and service points	4,862	97	6	26	7	11%	31%	28%	24%	5%
4. Medium socioeconomic level, high concentration of transportation points and highly residential	17,824	118	5	34	4	8%	29%	36%	25%	3%
5. Medium socioeconomic level, high concentration of transportation points and residential	6,771	109	4	25	2	10%	32%	31%	23%	4%
6. Medium socioeconomic level, somewhat residential with few points of interest	4,334	110	2	15	2	7%	23%	29%	33%	8%
7. Medium socioeconomic level, not very residential with very low population density	2,657	76	3	15	3	7%	24%	29%	32%	8%
8. Medium socioeconomic level, highly residential with very few points of interest	4,430	143	1	9	1	6%	26%	33%	29%	6%
9. Medium socioeconomic level, very high concentration of commerce, transportation and service points	4,871	91	42	35	13	11%	33%	29%	22%	4%
10. Medium-low socioeconomic level, pure residential	4,199	151	1	9	1	2%	15%	30%	42%	11%
11. Medium-low socioeconomic level, pure residential with low density	1,964	90	1	9	1	3%	16%	29%	41%	11%
12. Medium-low socioeconomic level, with high concentration of commerce, transportation and service points, low density	5,568	96	5	26	5	4%	18%	30%	38%	9%
13. Low socioeconomic level, residential with high population density	3,323	224	0	8	1	1%	7%	22%	52%	19%
14. Low socioeconomic level, residential with high population density, some concentration of transportation points	8,202	153	1	21	3	1%	6%	22%	54%	18%
15. Low socioeconomic level, residential with medium population density	5,028	164	1	12	2	1%	6%	21%	52%	19%
Total Population Averages	**4,771**	**133**	**4**	**16**	**2**	**10%**	**20%**	**25%**	**35%**	**10%**

The following descriptions summarize each cluster's main characteristics:

- Cluster 1, High socioeconomic level, high concentration of commerce, transportation and service points, low population density: dominated by high income households, these census tracts are located outside the city's main ring, in areas where there are few apartment buildings (people live mainly in houses). These areas, due to their distance from the city's downtown, usually have their own service and commerce concentrations, a few blocks away from the main population centers. Although the transportation index is high, this is explained by the presence of gas stations and avenues, not by subway or bus stops. People usually drive in, within and out these areas due to limited public transportation coverage.
- Cluster 2, Medium-high socioeconomic level, with concentration of commerce, transportation and service points: high income areas (although not as high as cluster 1), closer to the city's downtown. In these areas, people live in apartment buildings as well as houses. There is usually more coverage of public transportation, and people alternate its use with cars. These are areas usually marked by high pedestrian traffic.
- Cluster 3, Medium socioeconomic level, high concentration of transportation and service points: mid-income areas, with a high concentration of service and transportation points. High concentration of transportation points. These are probably residential areas with a mix of houses and apartment buildings, with service micro-centers and an extensive coverage of public transportation.
- Cluster 4, Medium socioeconomic level, high concentration of transportation points and highly residential: highly residential mid-income areas with a big coverage of public transportation. Pedestrian traffic mainly driven by reaching public transportation stops (not by commerce or service concentrations).
- Cluster 5, Medium socioeconomic level, high concentration of transportation points and residential: similar to cluster 4, but not as residential and with fewer transportation points.
- Cluster 6, Medium socioeconomic level, somewhat residential with few points of interest: mid-income areas not highly residential. Mostly peripheral tracts with a few apartment buildings, but not a lot of commerce, service or transportation points.
- Cluster 7, Medium socioeconomic level, not very residential with very low population density: like cluster 6, but with lower population density, indicating a higher prevalence of houses instead of apartment buildings.
- Cluster 8, Medium socioeconomic level, highly residential with very few points of interest: highly dense populated area, probably with a lot of apartment buildings.
- Cluster 9, Medium socioeconomic level, very high concentration of commerce, transportation and service points: the "downtown." Census tracks with some people living, mostly in apartment buildings, but whose main demand driver is the concentration of commerce and service. Intensive pedestrian traffic.
- Cluster 10, Medium-low socioeconomic level, pure residential: mid-low income areas, highly populated and almost exclusively residential. Only transportation points, which are the only drivers of pedestrian traffic. High concentration of apartment buildings.

- Cluster 11, Medium-low socioeconomic level, pure residential with low density: like cluster 11, but with more houses instead of apartment buildings.
- Cluster 12, Medium-low socioeconomic level, with high concentration of commerce, transportation and service points, low population density: house-based residential tracts of mid-low income households. High concentration of commerce (for a mid-low income area) and transportation points. Pedestrian traffic associated with commuting (to/from public transportation stops), plus some commerce micro-centers.
- Cluster 13, Low socioeconomic level, residential with high population density: highly populated residential low income areas. No service or commerce concentration and low concentration of transportation points. Domestic pedestrian traffic mostly.
- Cluster 14, Low socioeconomic level, residential with high population density, some concentration of transportation points: highly populated, house based, residential low income areas, with a high concentration of transportation points. A lot of pedestrian traffic probably associated with commuting to and from the area.
- Cluster 15, Low socioeconomic level, residential with medium population density: like cluster 15, but with lower concentration of transportation points, indicating a more domestic pedestrian traffic (as opposite to commute driven pedestrian traffic).

The following page shows a map of Santiago where every census tract has been shaded according to its cluster membership. While in some areas of the city there is a certain degree homogeneity (i.e. neighbor census tracts belonging to the same cluster), in others, contiguous census tracts were shaded in different colors. It is also possible to note that there are census tracts belonging to the same clusters in different and distant parts of the city.

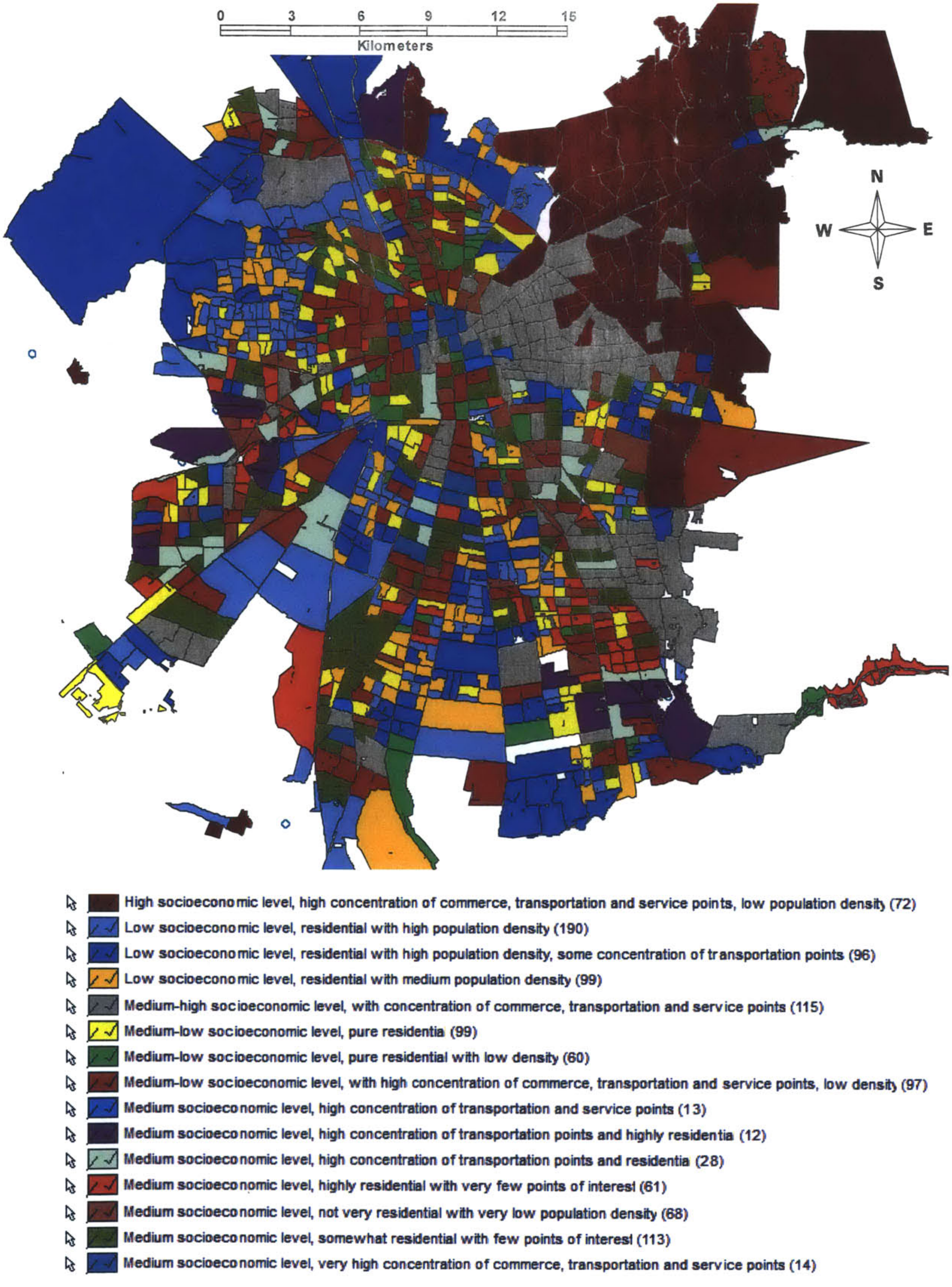

Figure 11: Census Tracts Colored by Cluster

Measuring "Relevant" Homogeneity

As was mentioned at the beginning of the chapter, the goal of the clustering process was to generate groups of census tracts with more homogeneous demand for the company's products. Homogeneity was measured in this case using the company's sales density, defined as follows, for every product category (instant games and lottery games):

$$\rho_{ij} = \frac{Lottery\ Sales_{ij}}{People_i \times Traffic\ Points_i} \left[\frac{\$}{Person \times Traffic\ Point} \right]$$

(For every census tract *i* and product category *j*).

In this regard, if the clustering process provided relevant homogeneity to the analysis, the overall variability of this index should decrease. To build this index, lottery sales were grouped, as mentioned, by product category, for the last three months of available data (data was available from September 2009 to December 2010, for lottery games, and from January 2010 to January 2011 for instant games).

Sales were not assigned to census tracts solely on the base of the stores' addresses. An address-only based assignation would lead to undesirable results as stores can also serve consumers from nearby census tracts. The pitfalls presented by a pure-address approach increase as stores get closer to the boundary between census tracts. The following figure presents an example of this situation, where two stores are located just at the edge of their census tracts, separated by an avenue:[7]

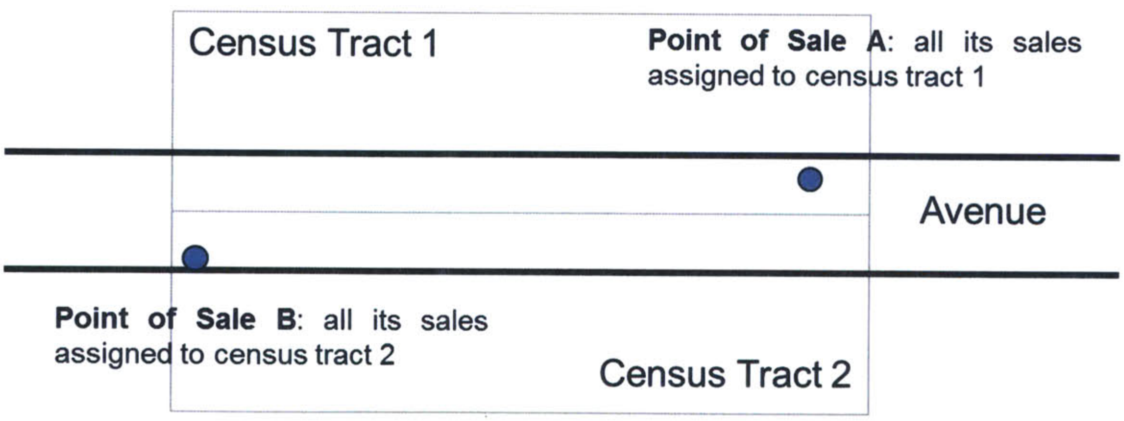

Figure 12: Example of an Undesirable Assignation

To deal with this issue, each store's sales were equally assigned to an influence radius around the store address. Based on experience of the author of this thesis, a 300 m radius was used, homogeneously assigning sales to the area formed by the circle. Circles were then intersected with the census tracts, assigning their sales according to the intersection proportion. The following map shows a real example of the intersection process:

[7] Avenues and streets are the typical criteria used to separate census tracts. In this regard, it is often the case at the boundary between census tracts that stores located across the street from each other belong to different census tracts.

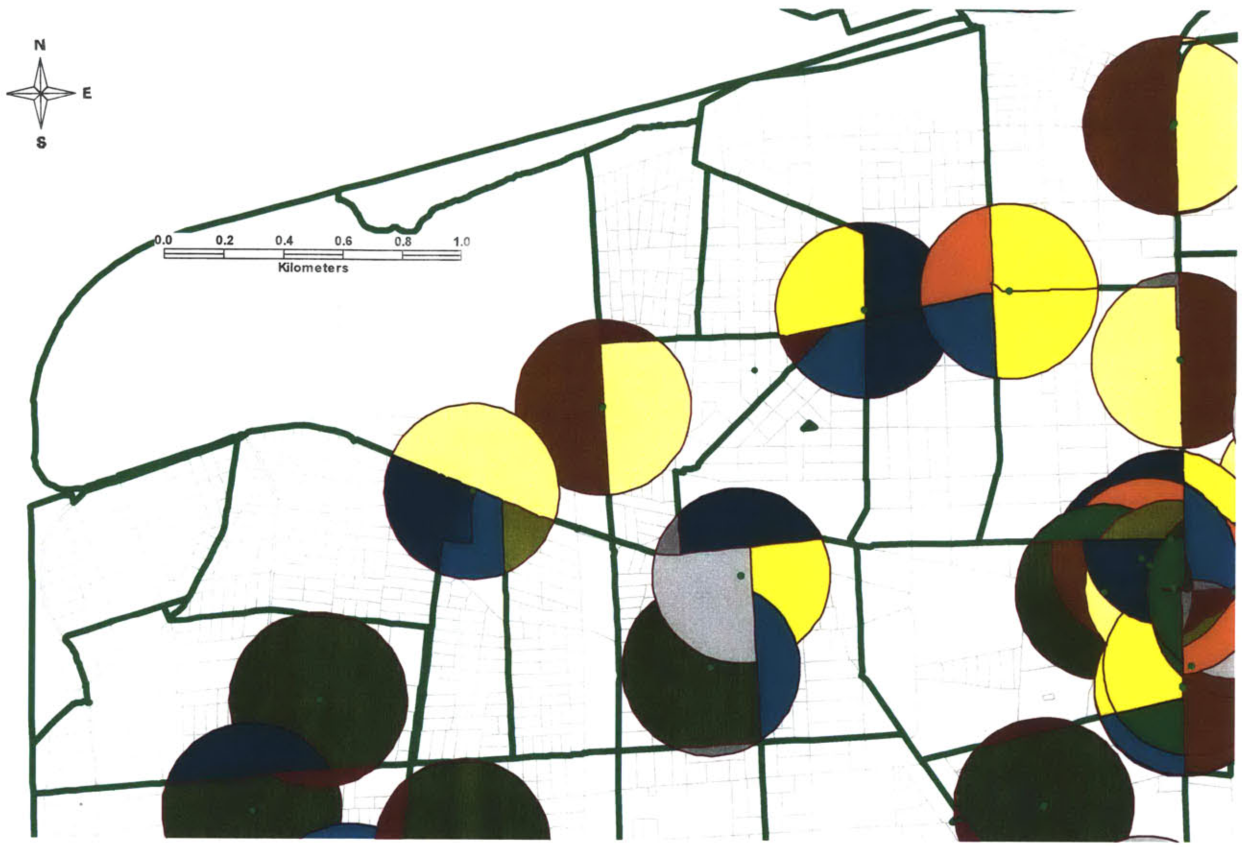

Figure 13: Intersection of Influence Area and Census Tracts

The thicker green lines represent census tract borders, the small green points at the center of the circles are stores, and the intersection between influence radiuses and census tracts are presented in different colors. As mentioned, sales were assigned to census tracts based on the proportion of their intersection with the influence radius.

The total area of the influence radius is constant (the area of a circle with a 300 m radius):

$$Area(Influence\ Radius) = \pi \times (300\ [m])^2 = 282{,}743.3\ [m^2]$$

If P_{zi} is the intersection between the influence radius around store *z* and census tract *i*, and S_{zj} the sales of the store *z* in the product category *j* (instant games and lottery games), then the total sales assigned to census tract *i* for product category *j* were:

$$S_{ij} = \sum_{z=1} \left(\frac{Area(P_{zi})}{282{,}743.3\ [m^2]} \times S_{zj} \right) [\$]$$

Thus, sales density for census tract *i* and product category *j* was:

$$\rho_{ij} = \frac{\sum_{z=1} \left(\frac{Area(P_{zi})}{282{,}743.3\ [m^2]} \times S_{zj} \right)}{People_i \times Traffic\ Points_i} \left[\frac{\$}{Person \times Traffic\ Point} \right]$$

The following table summarizes sales densities for both instant games and lottery games, before the clustering process described in this chapter:

Table 4: Sales Density before Clustering

	Sales Density	Instant Games	Lottery Games
	Number of Census Tracts	1,137	1,137
Sales Density $\left[\frac{\$}{Person \times Traffic\ Point}\right]$	Minimum	0	0
	Maximum	2,576	12,656
	Mean	44	190
	Std. Deviation	115.81	548.18
	Coefficient of Variation [dimensionless]	2.61	2.88

The following sales densities were obtained after the process, for each cluster and product category:

Table 5: Summary of Sales Density, Lottery Games, per Cluster

		Sales Density $\left[\frac{\$}{Person \times Traffic\ Point}\right]$				
Cluster	**Number of Census Tracts**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**	**Coefficient of Variation [dimensionless]**
1	72	0	1,231	177	188	1.07
2	115	0	1,294	226	236	1.04
3	13	8	242	138	68	0.49
4	12	19	68	39	18	0.46
5	28	12	240	89	60	0.68
6	113	0	1,226	161	182	1.13
7	68	5	12,656	442	1,524	3.45
8	61	0	1,011	195	205	1.05
9	14	26	1,394	325	391	1.20
10	99	0	573	148	124	0.84
11	60	0	8,655	424	1,131	2.66
12	97	3	963	164	181	1.10
13	190	0	7,957	195	623	3.20
14	96	0	383	52	67	1.28
15	99	0	1,515	100	177	1.77
		Coefficient of Variation's Weighted Average				**1.69**

Table 6: Summary of Sales Density, Lottery Games, per Cluster

Cluster	Number of Census Tracts	Sales Density $\left[\frac{\$}{Person \times Traffic\ Point}\right]$ Minimum	Maximum	Mean	Std. Deviation	Coefficient of Variation [dimensionless]
1	72	0	303	41	46	1.14
2	115	0	322	53	58	1.10
3	13	2	58	33	17	0.51
4	12	4	16	9	4	0.49
5	28	3	60	21	15	0.70
6	113	0	238	38	43	1.13
7	68	1	2,576	99	311	3.15
8	61	0	160	44	44	0.99
9	14	6	393	86	110	1.28
10	99	0	129	35	29	0.84
11	60	0	1,922	100	253	2.54
12	97	1	232	40	46	1.15
13	190	0	1,473	45	123	2.73
14	96	0	87	12	16	1.31
15	99	0	397	25	46	1.85
		Coefficient of Variation's Weighted Average				**1.61**

In both cases, sales densities' variability decreased considerably (measured by the coefficient of variation), indicating that the variables chosen for the clustering process added homogeneity to phenomena being studied. Sales densities' mean values also vary considerably across clusters, which is another indication of homogeneity within clusters and heterogeneity across them. The following table summarizes the results:

Table 7: Reduction in Heterogeneity after Demand Cluster

Product Category	Coefficient of Variation Before Clustering	Coefficient of Variation After Clustering
Instant Games	2.61	1.61
Lottery Games	2.88	1.69

Crossing Store Geoclusters with Demand Clusters

The purpose of this section is to characterize the demand faced by store geoclusters. In order to do so, store geoclusters were crossed with census tracts' demand-based clusters. This section describes the process followed.

Intersection of Store Geocluster and Census Tracts

After the geoclustering process was completed and nearby stores were grouped together (Chapter 2), the next step was to characterize the demand served by each store geocluster. This task was done by crossing store geoclusters created in the previous chapter with the census tracts cluster layer[8] (subsequently referred as "demand clusters"). Store geoclusters were then assigned demand clusters according to their intersection with census tracts.

The procedure followed was based on spatial intersection, due to the geographical overlapping between store geoclusters and census tracts. The following figure illustrates the first step of the process, when store geoclusters layers were placed on top of the demand cluster layer:

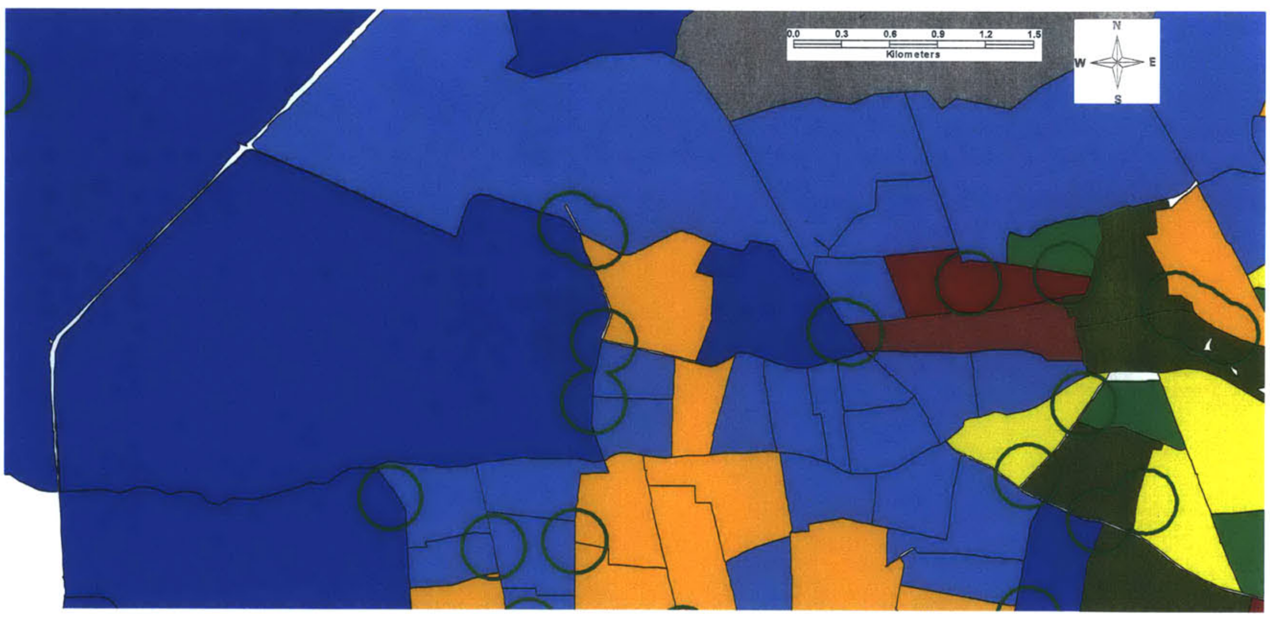

Figure 14: Overlapping Between Census Tracks and Store Geo Clusters

The shapes shaded with different colors are census tracts (colored according to their cluster membership, as shown at the beginning of this chapter), and the hollow circular forms with green borders are store geoclusters (the ones formed with 200 m radius are displayed in this illustration).

---

[8] So far, two clustering procedures have been described:

- At the beginning of this chapter, census tracts were clustered based on demand characteristics, to obtain groups of them facing more homogeneous demands.
- In Chapter 2, stores were clustered based on their location, to create group of nearby stores or "store neighborhoods."

The next step was to create elements of intersection between store geoclusters and census tracts, as the following figure illustrates:

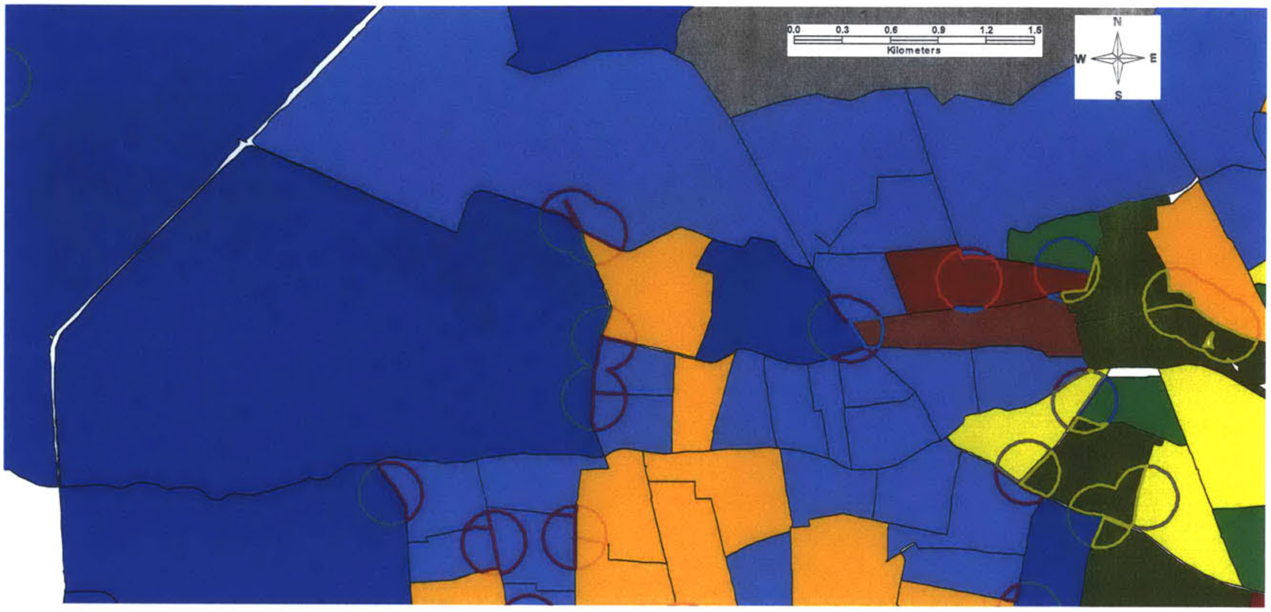

Figure 15: Intersection Elements, Store Geoclusters and Census Tracts

Elements of intersection are represented by hollow shapes, with borders colored differently to represent intersections with different clusters.

Grouping Back to Characterize Store Geoclusters

Using this new layer of elements, the next step was to analyze the profile of the elements of intersection in terms of the cluster (demand cluster) they intersect with, and then to group them back to rebuild the store geoclusters:

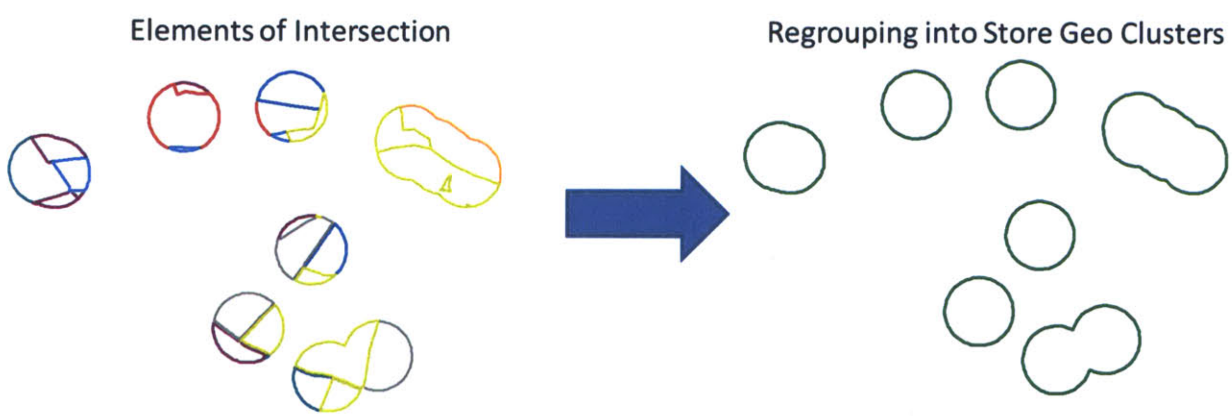

Figure 16: Regrouping of Elements of Intersection into Store Geoclusters

When grouping back the elements of intersection into store geoclusters, some of them contained elements of intersection classified into different demand clusters. To deal with this situation, the final allocation of demand clusters to store geoclusters was based on the percentage of the area intersected: store geoclusters were assigned the demand cluster with which they had the biggest proportion of its area intersected (subsequently referred as "dominant demand cluster"). A potential problem associated with this procedure was that, in some cases, even the demand cluster accounting for the biggest proportion of a store geocluster's area could still represent a rather small proportion (for example, if the dominant demand cluster represents 30% of the area, and the second 25%; cases like this are more frequently observed in store geoclusters covering bigger areas, characterized by several demand profiles). To characterize the severity of the problem, the distribution of the areas accounted by the dominant demand clusters was analyzed. The following table summarizes the results:

Table 8: Distribution of the Area Intersected by the Dominant Demand Cluster by Geocluster Radius

Geoclusters	Less than 40%	More than 40% and less than 100%	Pure Clusters (100%)
150 m	4%	61%	35%
200 m	6%	71%	23%
300 m	13%	71%	16%

As shown in the table, for the store geoclusters configured using 150 m radiuses for example, only in 4% of the cases did the dominant demand cluster represent less than 40% of the area. As expected, as the size of the radius of the geoclusters increases, the number of cases where the dominant demand cluster represents less than 40% of the area also increases (by covering a bigger area, the likelihood of intersection with more demand clusters grows). In any case, however, in more than 87% of the store geoclusters, the dominant demand cluster represented more than 40% of the area. In cases where less than 40% of the area was accounted by the dominant demand cluster, the second and third demand clusters in importance were frequently very "close" (in terms the variables used to create the demand clusters) to the dominant one, which indicates that misclassification risks were low.

Chapter 4: Cannibalization Analysis

The purpose of this chapter is to analyze the information gathered so far, looking for evidence of cannibalization in historic data, setting the grounds for the location analysis presented in the next chapter.

If cannibalization existed, proving that demand is not "infinite" and that every new store incorporated to the distribution network caused nearby stores to decrease their sales, the way to measure it would be by finding the relationship between changes in the amount of active (i.e., with transactions) stores and their average sales. This relationship, however, depends on a series of other variables that needed to be controlled:

- The demand profile of the area served by the stores in terms of its sales potential. The hypothesis is that areas with higher potential should be able to support more stores without showing serious signs of cannibalization.
- The average distance between stores. The hypothesis is that as the average separation between stores grows, the severity of the cannibalization should be lower.
- The type of product. The hypothesis is that products whose purchase decision is mostly driven by impulse would be subject to less cannibalization than products with a more planned purchase decision.

To deal with the previous hypotheses, two clusters analyses were implemented:

- In Chapter 2, geographic clusters were built grouping "nearby" stores, creating store neighborhoods or geoclusters. Buffer distances of 150, 200 and 300 m were used. This is related to the second hypothesis presented. Each of the store geoclusters was then assigned a demand cluster. Cannibalization analysis was then performed within each of these geoclusters, looking for the relationship between variations in the clusters' stores density and the average size of the stores.
- In Chapter 3, a cluster analysis considering variables related to the demand potential was implemented, and groups ("demand clusters") of census tracts with similar demand profile were created. Cannibalization analysis was then performed within each of the clusters created. This is related to the first hypothesis presented at the beginning.

To deal with the third hypothesis, the analysis was broken down by product category:

- One for instant games, whose purchase process is mainly driven by impulse.
- Another for lottery games, whose purchase is more planned and less driven by impulse.

There was still another consideration that needed to be taken into account: seasonality. This is a highly seasonal industry, but not in the traditional sense of, for example, cold versus hot weather or school versus vacation time. In this industry, variations in demand among different times during the year are driven almost exclusively by the accumulation jackpots: when the jackpot accumulates week after week, demand rises as the prize becomes more attractive.[9] In this regard, after the jackpot is won by one (or many) bettor(s), demand decreases and the average sales per store falls. This decrease in sales, however, is not explained by the presence of cannibalization, and models built to measure it should account for this effect. Seasonality was expected to be more important for lottery than for instant games.

Building the Variables

The first step of the process was to assign transactions to each of the geoclusters created in Chapter 2. This process was done by using the stores located inside the geoclusters, for each of the three different layers (150, 200 and 300 m): each geocluster "inherited" the transactions of the stores generating it. Transactions were accumulated on a monthly basis, so what was obtained was a database for each of the geocluster layers with sales accumulated by month and product category (lottery and instant games). Along with the accumulated sales, two other variables were calculated for each month and geocluster:

- The active number of stores per product category, i.e., the number of stores with at least one transaction for the product category during the month. The hypothesis was that variations in this indicator were related to variations in the average sales per store within the geoclusters.
- The average sales per store per product category. This was calculated as the sum of all the transactions for all the active stores, per product category and month divided by the number of active stores per product category during each month. This number gives an indication of the average sales of each active store; a decrease in this number, therefore, coupled with an increase in the number of active stores, would evidence the presence of cannibalization.

The following table and figure show an example of how this database was assembled. Geocluster 5, from the 300 m radius layer, includes transactions from 14 stores:

[9] Demand increases in these periods as (1) customers who usually do not buy tickets enter the market just for these highly attractive prizes and (2) regular customers increase their average spending.

Figure 17: Geocluster 5, 300 m Radius Layer

Based on the information from the stores' transactions, the following set of records for the working database was obtained (instant games example):

Table 9: Geocluster 5, 300 m Radius Layer, Instant Games, Records on the Database

Geocluster Id	Month	Active Stores	Sum of sales [$]
5	01-Jan-10	7	1,421,978
5	01-Feb-10	7	267,007
5	01-Mar-10	7	707,124
5	01-Apr-10	6	1,328,197
5	01-May-10	6	1,827,707
5	01-Jun-10	7	1,706,714
5	01-Jul-10	7	1,678,640
5	01-Aug-10	8	1,580,983
5	01-Sep-10	8	1,063,054
5	01-Oct-10	10	2,526,704
5	01-Nov-10	11	2,220,435
5	01-Dec-10	11	2,538,586
5	01-Jan-11	11	2,041,821

As seen, of a universe of potentially 14 different stores, the highest number of stores that were simultaneously active was 11.

To measure cannibalization, as mentioned before, it is necessary to find the relationship between variations in the average size of stores and the number of active stores. However, these variables cannot be directly used because geoclusters have different numbers of active stores: there are, for example, some geoclusters with only one active store, while there are others with more than 50. In the first case, one additional store means an increment of 100% in the number of stores, while in the latter, of only 2%. A similar situation occurs with the average size of the stores. Accordingly, it is not advisable to build a model considering absolute variations between these two variables; the only way to do so would be to build one model per geocluster, which would reduce the number of observations available to at most the total number of months of data available, rendering any statistical analysis almost certainly non-significant. To avoid this issue, all variables were normalized by the averages of the geocluster for each of them. The next table follows the example for geocluster 5:

Table 10: Geocluster 5, 300 m Radius Layer, Instant Games, Records on the Database with Normalized Variables

Geocluster Id	Month	Active Stores	Sum of sales [$]	Average Size [$]	Active Stores Normalized	Average Size Normalized
5	01-Jan-10	7	1,421,978	203,140	0.86	1.034
5	01-Feb-10	7	267,007	38,144	0.86	0.194
5	01-Mar-10	7	707,124	101,018	0.86	0.514
5	01-Apr-10	6	1,328,197	221,366	0.74	1.127
5	01-May-10	6	1,827,707	304,618	0.74	1.551
5	01-Jun-10	7	1,706,714	243,816	0.86	1.241
5	01-Jul-10	7	1,678,640	239,806	0.86	1.221
5	01-Aug-10	8	1,580,983	197,623	0.98	1.006
5	01-Sep-10	8	1,063,054	132,882	0.98	0.677
5	01-Oct-10	10	2,526,704	252,670	1.23	1.286
5	01-Nov-10	11	2,220,435	201,858	1.35	1.028
5	01-Dec-10	11	2,538,586	230,781	1.35	1.175
5	01-Jan-11	11	2,041,821	185,620	1.35	0.945
	Averages	**8.15**		**196,411**		

As mentioned before, there was another effect that had to be controlled: seasonality. As suggested, seasonality in this industry is more related to jackpot accumulation than to weather or other external factors, and it was expected to play a more important role in lottery games than in instant games. In this regard, seasonal factors for each month and product category were calculated by dividing the average sales of the company by the product category in the month by the average sales of the company for the product category for the entire period. The following figure shows lottery games' seasonal factors, for the period with available information:

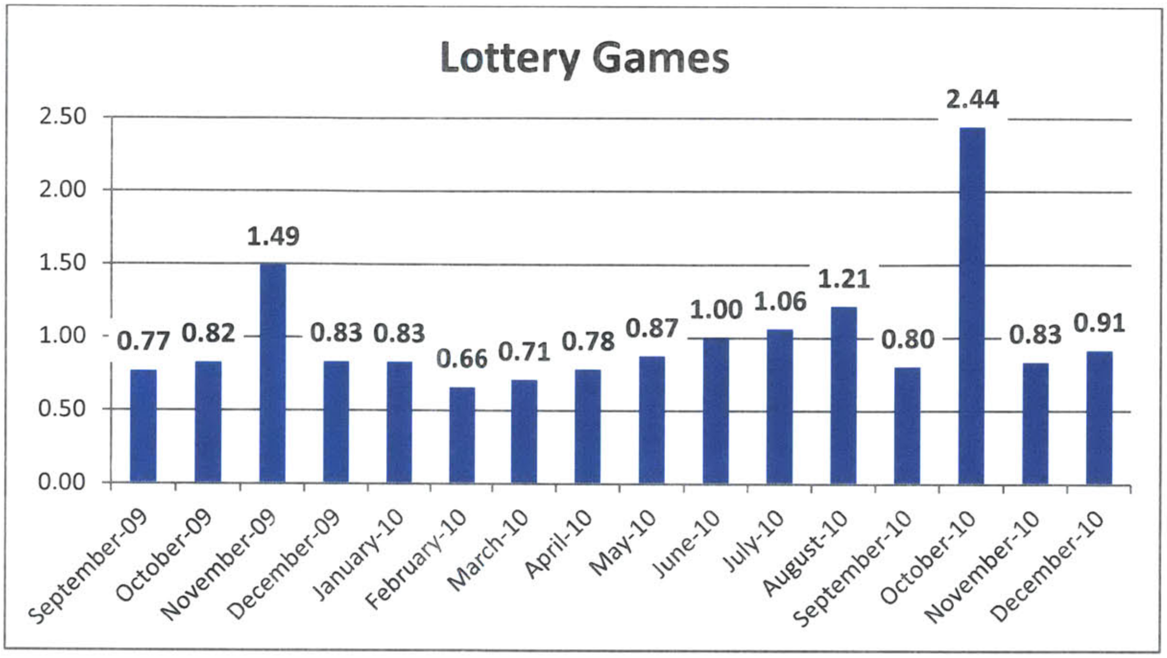

Figure 18: Seasonal Factors, Lottery Games

As shown in the graph, there were two months with high seasonal factors: November 2009 and October 2010. Both were months with a great accumulation of jackpots.

The following figure shows instant games' seasonal factors, for the period with available information:

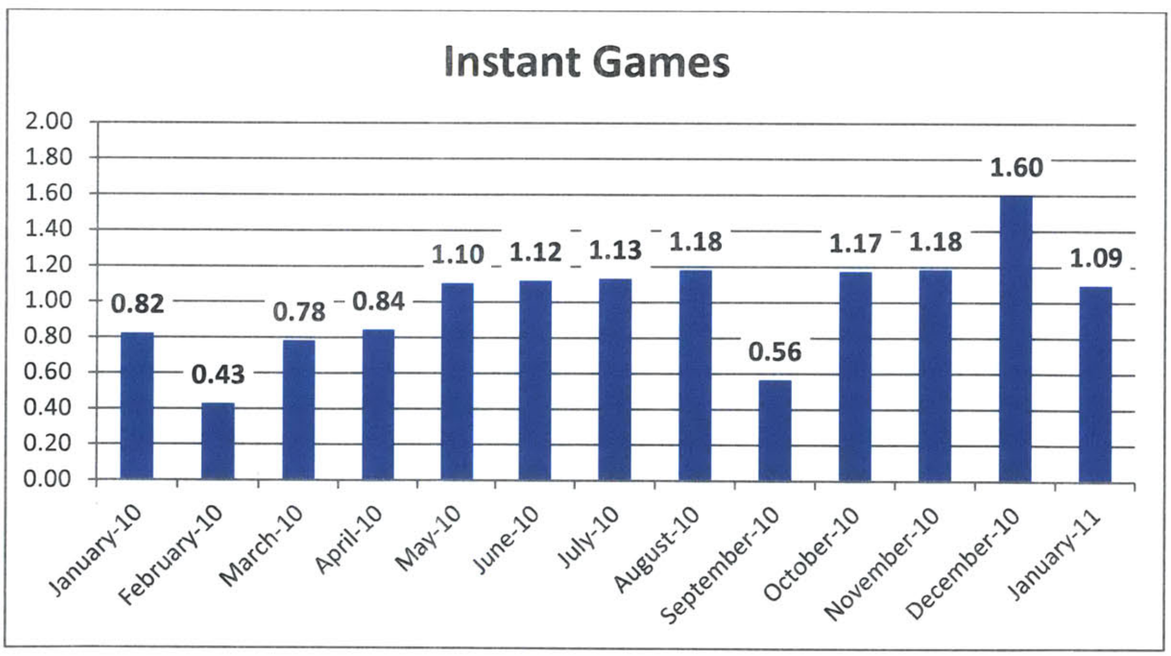

Figure 19: Seasonal Factors, Instant Games

Interestingly enough, instant games' seasonal factors, although not as marked as in the case of lottery games, showed notable variations among months. February and September are months in which people usually take vacations, leaving the city for which this analysis was conducted (Santiago), and in December, because of the holidays, people tend to buy more games due to (1) the appearance of special editions and (2) the opportunity to use them as gifts for Christmas.

The seasonal factors previously shown were calculated using the sales of all the stores. In following sections, seasonal factors were calculated for each demand cluster, recognizing that seasonality may vary as the demand profile of the consumers varies.

In summary, the following variables were built and subsequently used in the analysis:

T_{ij} : Average size of the stores of the geo cluster *j* at the time *i*

A_{ij} : Number of active stores in the geo cluster *j* at the time *i*

$\overline{T_j}$: Average size of the stores of the geo cluster *j*

$\overline{A_j}$: Average number of active stores at the geo cluster j

$t_{ij} = \frac{T_{ij}}{\overline{T_j}}$: Normalized size of the stores of the geo cluster j at the time i

$a_{ij} = \frac{A_{ij}}{\overline{A_j}}$: Normalized number of active stores at the geo cluster j at the time i

f_i : Seasonal factor for time *i*

These variables were built for every product category (instant and lottery games), and for each geoclustering layer (150, 200 and 300 m). The following table continues the example of geocluster 5 for the 300 m layer, for instant games:

Table 11: Geocluster 5, 300 m Radius Layer, Instant Games, Records on the Database with Normalized Variables and Seasonal Factors

j	i	A_{ij}		T_{ij}	$a_{ij} = \frac{A_{ij}}{\overline{A_j}}$	$t_{ij} = \frac{T_{ij}}{\overline{T_j}}$	f_i
Geo Cluster Id	**Month**	**Active Stores**	**Sum of sales [$]**	**Average Size [$]**	**Active Stores Normalized**	**Average Size Normalized**	**Seasonal factor**
5	1-Jan-10	7	1,421,978	203,140	0.86	1.03	0.82
5	1-Feb-10	7	267,007	38,144	0.86	0.19	0.43
5	1-Mar-10	7	707,124	101,018	0.86	0.51	0.78
5	1-Apr-10	6	1,328,197	221,366	0.74	1.13	0.84
5	1-May-10	6	1,827,707	304,618	0.74	1.55	1.10
5	1-Jun-10	7	1,706,714	243,816	0.86	1.24	1.12
5	1-Jul-10	7	1,678,640	239,806	0.86	1.22	1.13
5	1-Aug-10	8	1,580,983	197,623	0.98	1.01	1.18
5	1-Sep-10	8	1,063,054	132,882	0.98	0.68	0.56
5	1-Oct-10	10	2,526,704	252,670	1.23	1.29	1.17
5	1-Nov-10	11	2,220,435	201,858	1.35	1.03	1.18
5	1-Dec-10	11	2,538,586	230,781	1.35	1.18	1.60
5	1-Jan-11	11	2,041,821	185,620	1.35	0.95	1.09
	Averages	**8.15**		**196,411**			
		$\overline{A_j}$		$\overline{T_j}$			

As expected, the distribution of the new variable a_{ij} is highly concentrated around 1 (and the frequency of "1" as an observation is also very high). The following were the distributions obtained for both product categories and every geocluster layer (150, 200 and 300 m) for this variable:

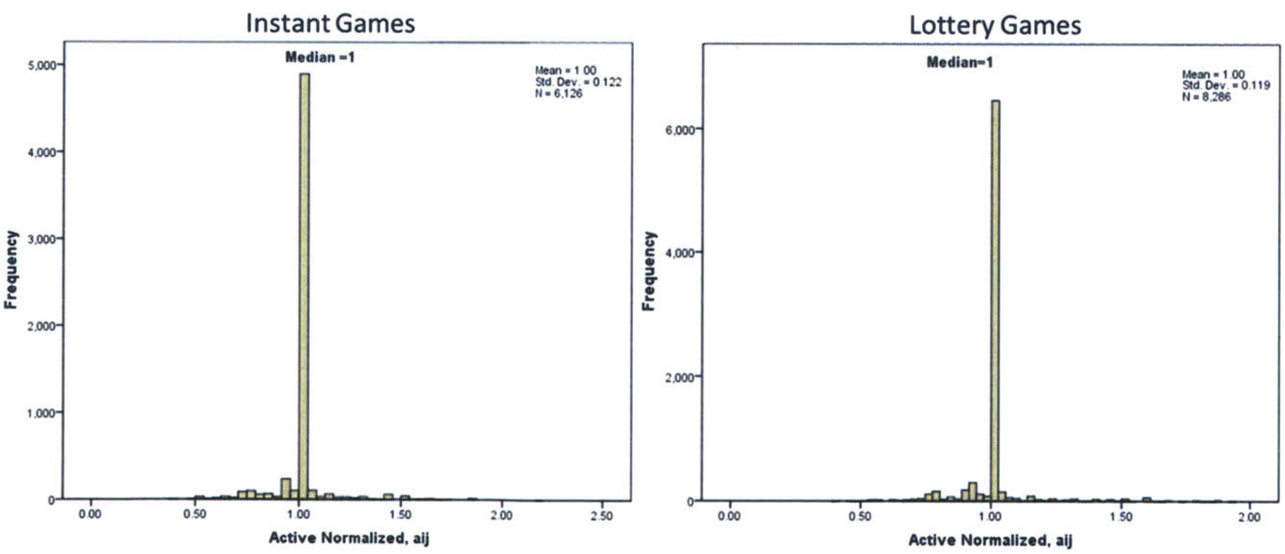

Figure 20: Distribution for Active-Normalized, 150 m Geocluster Layer

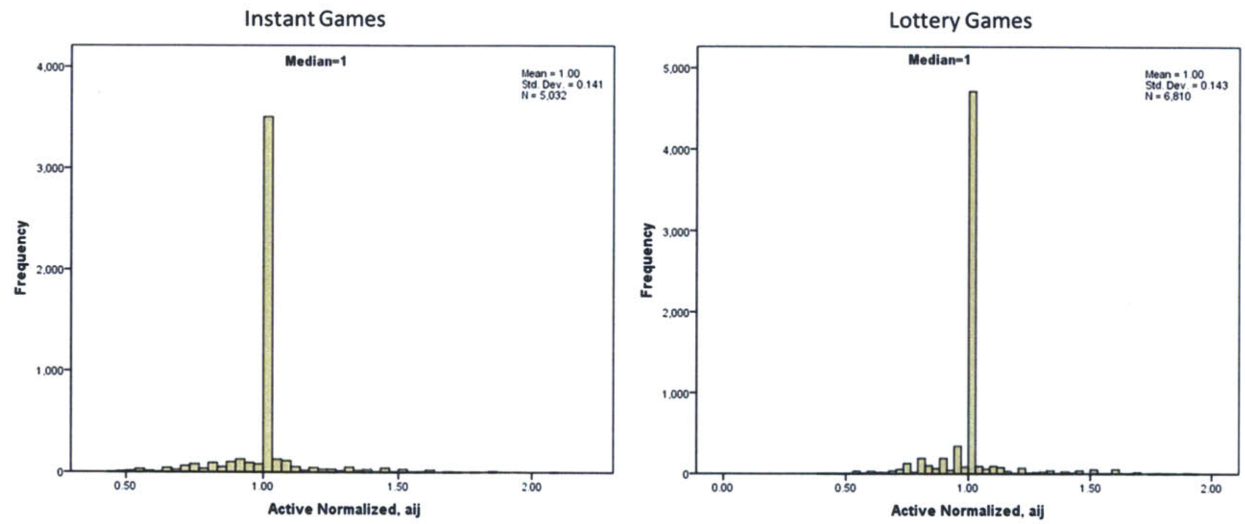

Figure 21: Distribution for Active-Normalized, 200 m Geocluster Layer

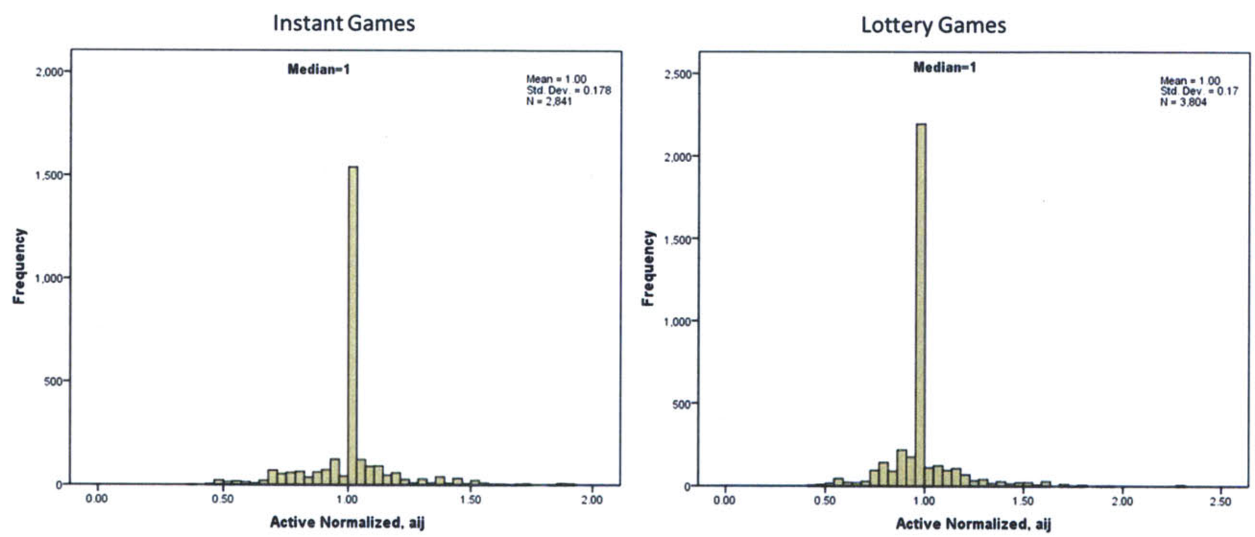

Figure 22: Distribution for Active-Normalized, 300 m Geocluster Layer

As seen in the distribution graphs, the frequency of values close to 1 (and of "1" itself) was very high. This is explained by the presence of uni-store geoclusters, as mentioned in Chapter 3. What is interesting to note, however, is that the relative height of the bar representing values close to 1 decreases as the radius increases. This situation was also expected based on what was described in Chapter 3: as the size of the radius increases, the chance of getting geoclusters with more than one store also increases, and geoclusters with more stores are more likely to present variations during the period of observation in the number of active stores. These variations were the data source for the cannibalization analysis. The diminution in the relative height of bars closer to 1 is captured by the kurtosis measurement, which decreases as the radius increases:

Table 12: Kurtosis for Active-Normalized

Product	150 m	200 m	300 m
Instant Games	14.853	8.095	4.658
Lottery Games	17.279	9.831	9.456

The distribution for the other normalized variable, t_{ij}, was not so concentrated around 1, although it still had a higher frequency of cases near the mean compared to a normal distribution with similar parameters, particularly for lottery games. The following figures show the distributions:

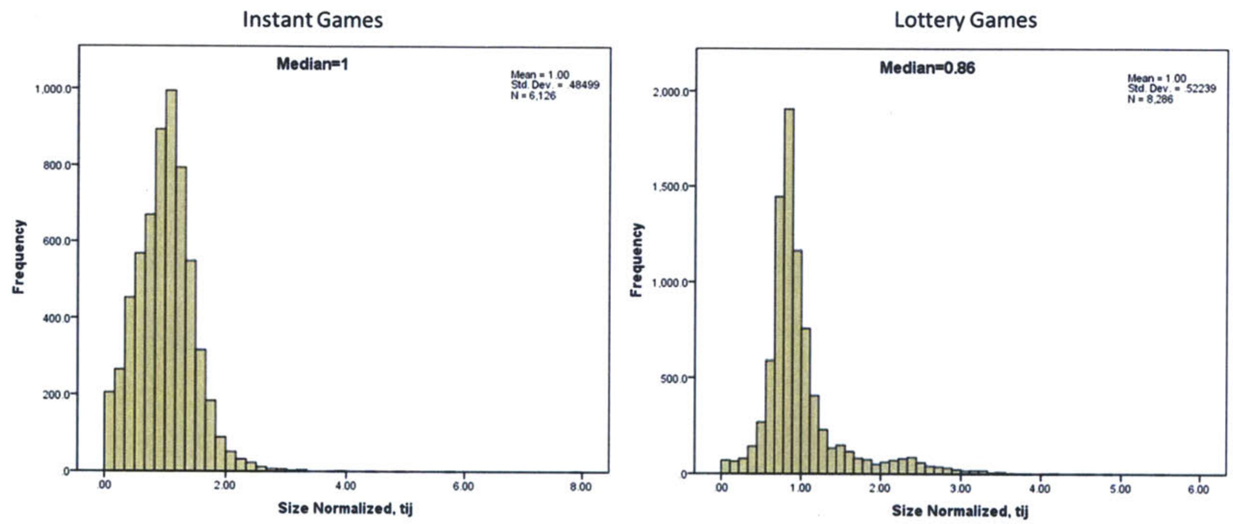

Figure 23: Distribution for Size-Normalized, 150 m Geocluster Layer

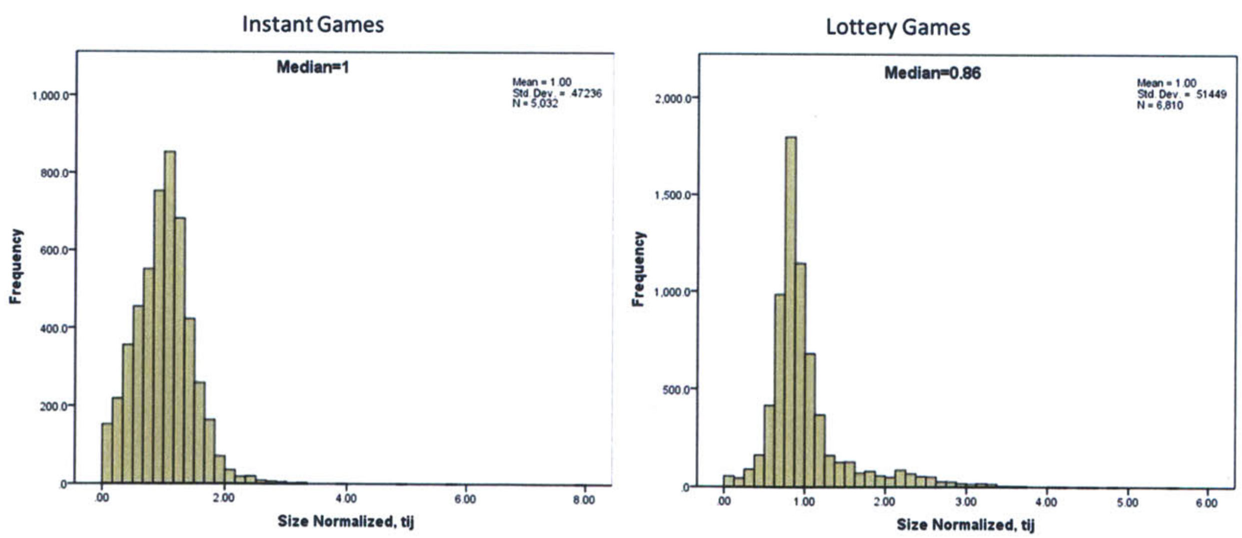

Figure 24: Distribution for Size-Normalized, 200 m Geocluster Layer

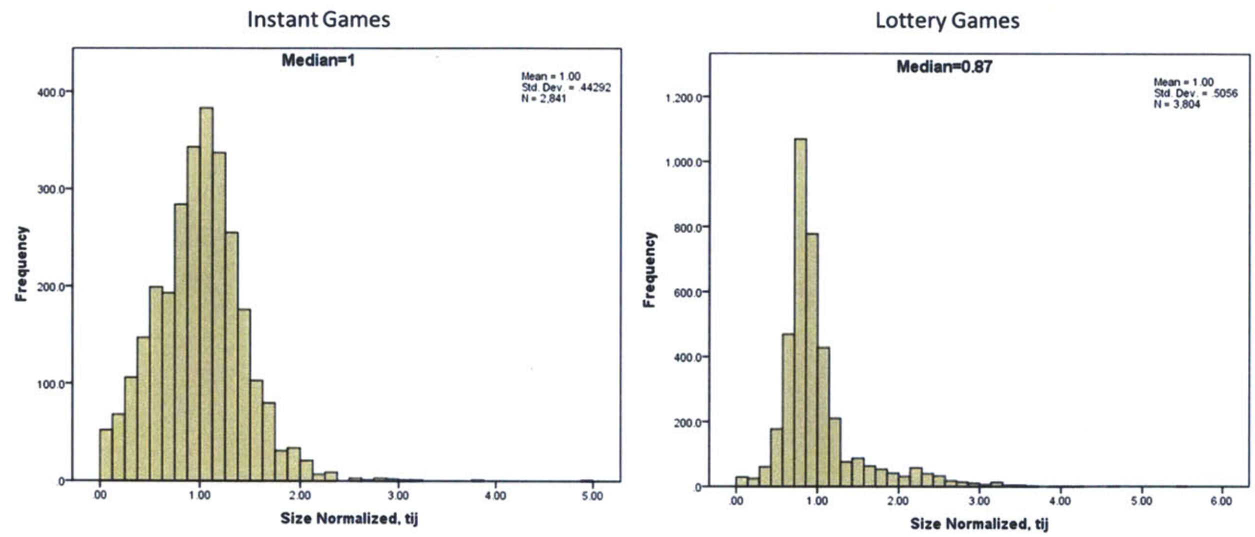

Figure 25: Distribution for Size-Normalized, 200 m Geocluster Layer

Modeling Aggregated Cannibalization

Once all the variables were built, the next step was to select the models (i.e., functional forms) to use. In this regard, the following three were originally considered:

- A linear model: $t_{ij} = \alpha \cdot a_{ij} + \beta \cdot f_i + C$
- An exponential model: $t_{ij} = C \cdot {a_{ij}}^{\alpha} \cdot {f^{\beta}}_i$
- A differences model: $\Delta t_{ij} = \alpha \cdot \Delta a_{ij} + \beta \cdot \Delta f_i + C$ were $\Delta X = X_{t-1} - X_t$

In the models, α is the coefficient representing cannibalization, while β is related to the seasonality effects.

In almost every test executed, the linear and differences models outperformed the exponential models; for this reason, only the results of these two models are subsequently shown. This subsection reports on the results obtained at an aggregated level; a subsequent section reports results at the demand cluster level.

Least square regression was employed to estimate the coefficients, and bootstraps with 500 samples were used to assess the robustness of the estimations. This section reports, as mentioned, individual results for every layer of geocluster (150, 200 and 300 m) for every product category (instant and lottery games).

Lottery Games

For lottery games, the following results were obtained for the linear model:

Table 13: Summary of Results, Aggregated Analysis, Lottery Games, Linear Model

Radius [m]	R	R^2	Adjusted R^2	Std. Error of the Estimate	F test	F test prob.
150	0.88	0.78	0.78	0.25	14663.28	0.00
200	0.89	0.79	0.79	0.24	12813.90	0.00
300	0.90	0.80	0.80	0.23	7608.78	0.00

As seen, for all layers the regressions gave significant results, with high Adjusted R^2's and an overall significance test (F-test) rejecting the hypothesis that coefficients were zero. Adjusted R^2's increases slightly as the radius increases, which may be a consequence of the presence of additional geoclusters with more than one store in them as the radius grows. The following is the distribution of the standardized residuals obtained for each layer:

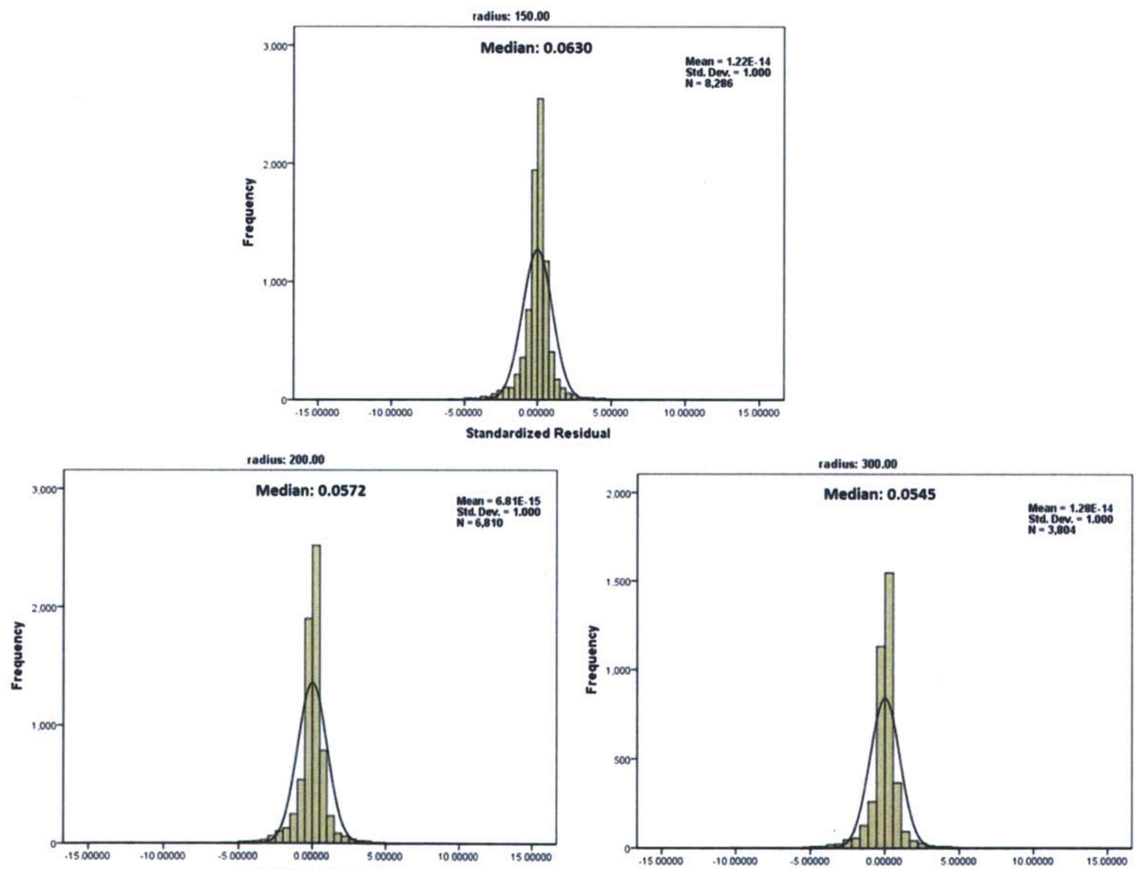

Figure 26: Standardized Residuals, Lottery Games, Linear Model

In every case, the standardized residuals showed a distribution highly concentrated around zero. Regressions were also run using the predicted values as dependent variables and the residuals as independent. In every case, these regressions showed non-significant results (both R^2 and F-test values were zero in every case), confirming the significance of the original regressions.

The following table summarizes the coefficients obtained, along with the results of the bootstrap:

Table 14: Summary of Results with Coefficients, Aggregated Analysis, Lottery Games, Linear Model

		Constant			Active Normalized (cannibalization coefficient)			Seasonal Factor		
	Radius [m]	**150**	**200**	**300**	**150**	**200**	**300**	**150**	**200**	**300**
Regression	Unstandardized Coefficients	0.541	0.579	0.641	-0.547	-0.583	-0.638	0.989	0.988	0.983
	Std. Error	0.023	0.021	0.022	0.023	0.02	0.022	0.006	0.006	0.008
	Standardized Coefficients				-0.124	-0.162	-0.215	0.882	0.888	0.899
	t	23.132	27.756	28.497	-23.887	-28.66	-29.20	169.38	157.36	122.24
	Sig.	0	0	0	0	0	0	0	0	0
Bootstrap	Bias	-0.001	-0.003	2.1E-5	0.001	0.001	-0.001	0	0.001	0.001
	Std. Error	0.033	0.028	0.031	0.032	0.026	0.028	0.012	0.013	0.018
	Sig. (2-tailed)	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	95% interval, lower	0.477	0.523	0.578	-0.611	-0.63	-0.695	0.965	0.964	0.951
	95% interval, Upper	0.602	0.632	0.702	-0.481	-0.533	-0.579	1.011	1.014	1.021

As seen in the table, the coefficients for Active Normalized and Seasonal Factor were significant, and the bootstrap shows that their estimation seems to be robust, as the 95% confidence intervals are relatively small around the values estimated using squares. From this analysis, the following conclusions can be drawn:

- The signs of the coefficients are aligned with what was expected:
    - Positive coefficients for seasonal factors, which indicate that stores increase their sales when the season is high (basically, accumulation of jackpots).
    - Negative coefficients for the number of active nearby stores (Active Normalized). This is a very important finding as it confirms the presence of cannibalization, which means that demand for this product category is not infinite (or large enough compared to the current sales of the distribution network), and that there are already sings of saturation, because increases in the number of active stores have effectively had an effect on the sales of nearby stores. This is one of the main hypotheses of this thesis and the results so far seem to confirm it for this product category. The implications for the company's network growth strategy will be analyzed in the following chapter.
- Although relatively small, there seems to be a correlation between the size of the cannibalization coefficient and the radius. As the radius increases, the coefficients tend to become more negative, indicating a slight increase in the effects of cannibalization. The bootstrap intervals, however, overlap, suggesting that these apparent differences may not be significant, and may be the results of just more observations within each geocluster with

changes in the number of active stores (as the size of the geocluster increases, so does the number of average stores per geocluster, which increases the likelihood of having more changes in actives stores, giving more cases to the calculation procedure to estimate the effect).

For the differences model, the following results were obtained:

Table 15: Summary of Results, Aggregated Analysis, Lottery Games, Differences Model

Radius [m]	R	R^2	Adjusted R^2	Std. Error of the Estimate	F test	F test prob.
150	0.938	0.88	0.88	0.2780768	27988.376	0
200	0.94	0.884	0.884	0.2692302	24061.568	0
300	0.942	0.887	0.887	0.2628285	13857.199	0

As seen, the differences model delivered slightly better results for every case (in the liner model, the Adjusted R^2's were in the proximity of 0.8 whereas in this case they are closer to 0.89), with, again, all the regressions producing significant results. The following was the distribution of the standardized residuals obtained for each layer:

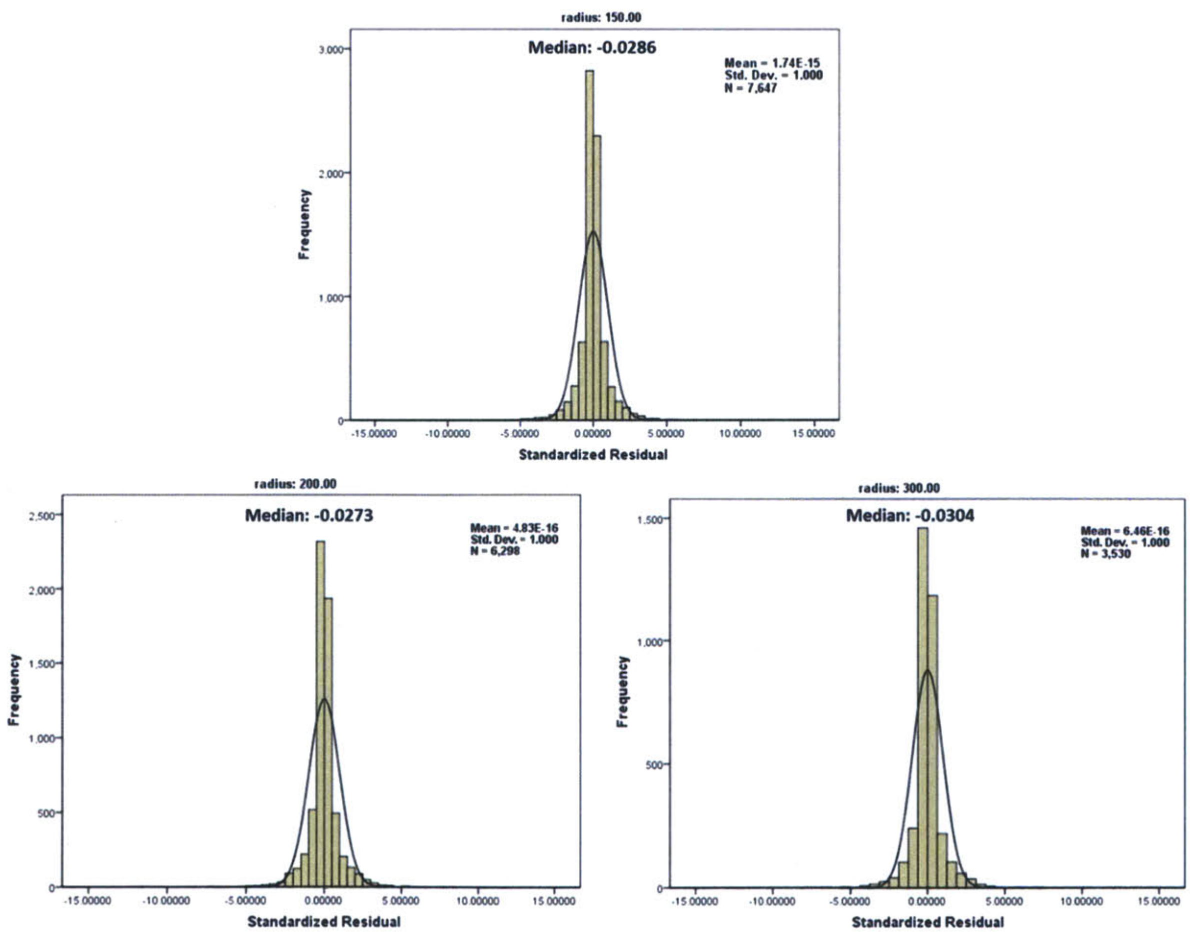

Figure 27: Standardized Residuals, Lottery Games, Differences Model

In every case, the standardized residuals showed a distribution highly concentrated around zero. Regressions were also run using the predicted values as dependent variables and the residuals as independent. In every case, these regressions again showed non-significant results (both R^2 and F-test values were zero in every case), confirming the significance of the original regressions.

The following table summarizes the coefficients obtained, along with the results of the bootstrap:

Table 16: Summary of Results with Coefficients, Aggregated Analysis, Lottery Games, Differences Model

		Constant			Active Normalized (cannibalization coefficient)			Seasonal Factor		
	Radius [m]	**150**	**200**	**300**	**150**	**200**	**300**	**150**	**200**	**300**
Regression	Unstandardized Coefficients	0.009	0.008	0.010	-0.622	-0.636	-0.757	1.084	1.076	1.065
	Std. Error	0.003	0.003	0.004	0.031	0.030	0.039	0.005	0.005	0.006
	Standardized Coefficients				-0.080	-0.089	-0.111	0.933	0.934	0.933
	t	2.672	2.487	2.185	-20.175	-20.856	-19.570	235.261	217.791	164.872
	Sig.	0.008	0.013	0.029	0.000	0.000	0.000	0.000	0.000	0.000
Bootstrap	Bias	0.000	0.000	0.000	0.000	0.000	-0.002	0.000	0.000	0.001
	Std. Error	0.003	0.003	0.004	0.038	0.039	0.051	0.008	0.009	0.012
	Sig. (2-tailed)	0.006	0.014	0.030	0.002	0.002	0.002	0.002	0.002	0.002
	95% interval, lower	0.003	0.001	0.001	-0.700	-0.708	-0.863	1.068	1.058	1.043
	95% interval, Upper	0.015	0.015	0.019	-0.548	-0.557	-0.665	1.100	1.095	1.090

As seen in the table, the coefficients for Active Normalized and Seasonal Factor were significant, and the bootstrap shows that their estimation seems to be robust, as the 95% confidence intervals are relatively small around the values estimated by least squares. From this analysis, conclusions similar to those in the case of the linear model can be drawn:

- The signs of the coefficients are aligned with what was expected:
  - Positive coefficients for the seasonal factors, which indicate that stores increase their sales when the season is high (basically, accumulation of jackpots). This is aligned again with what was expected.
  - Negative coefficients for the number of active nearby stores (Active Normalized). This, again, is a very important finding as it confirms the presence of cannibalization, indicating, as mentioned before, that there are already some signs of saturation in the distribution network, as increases in the number of active stores have effectively had an effect on the sales of nearby stores.
- Although relatively small, there seems again to be a correlation between the size of the cannibalization coefficient and the radius. As the radius increases, coefficients tend to become more negative, indicating a slight increase in the effects of cannibalization. The bootstrap intervals, however, overlap, suggesting that these apparent differences may not be significant, and may be the results of just more observations within each geocluster with changes in the number of active stores (as the size of the geocluster increases, so does the number of average

stores per geocluster, which increases the likelihood of having more changes in active stores, giving more cases to the calculation procedure to estimate the effect).

Both models, linear and differences, delivered results that were aligned with logic (in terms of magnitude and signs of the coefficients) and statistically significant, which seems to confirm the presence of cannibalization, evidencing the first signs of network saturation, as increases in the number of stores have had a negative effect on the average sales of active stores. The differences model gave slightly higher Adjusted R^2's, although in both cases (linear and differences model) Adjusted R^2's were greater than 0.8, indicating that regressions offered a good fit.

Although it was mentioned that exponential models were also tested, but consistently delivered worse results than the linear and differences models, they allow a direct interpretation of their coefficients as elasticities. In this regard, and only for illustrative purposes (as they were not used in subsequent analysis), their coefficients are shown:[10]

Table 17: Cannibalization Elasticities, Lottery Games

	Constant			Active Normalized (cannibalization coefficient)			Seasonal Factor		
Radius	**150**	**200**	**300**	**150**	**200**	**300**	**150**	**200**	**300**
Unstandardized Coefficients	-0.073	-0.068	-0.065	-0.541	-0.561	-0.612	1.045	1.047	1.041
Std. Error	0.005	0.005	0.007	0.043	0.038	0.039	0.015	0.016	0.020
Standardized Coefficients				-0.109	-0.141	-0.194	0.609	0.626	0.642
t	-14.933	-13.223	-9.926	-12.508	-14.911	-15.530	69.852	65.989	51.386
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

As mentioned, coefficients in the exponential model can be directly interpreted as elasticities: A coefficient of -0.541 for the case of the 150 m radius, for example, indicates that an increase of 1% in the concentration of active stores is expected to cause a decrease of 0.541% in the average size of the stores. In this regard, these coefficients provide a more intuitive interpretation.

A subsequent section in this chapter will provide results at the level of the demand clusters calculated in Chapter 2.

[10] Adjusted R^2's for the 150, 200 and 300 m radius were 0.373, 0.394 and 0.416 respectively.

Instant Games

For instant games, the following results were obtained for the linear model:

Table 18: Summary of Results, Aggregated Analysis, Instant Games, Linear Model

Radius [m]	R	R^2	Adjusted R^2	Std. Error of the Estimate	F test	F test prob.
150	0.523	0.273	0.273	0.41354	1150.737	0
200	0.529	0.280	0.279	0.40097	976.571	0
300	0.555	0.308	0.307	0.36863	631.019	0

Although still significant, Adjusted R^2's obtained in this case were substantially lower than what was observed with lottery games. These results, however, were expected since (1) customer's purchase decision associated with this product category is more driven by impulse, which weakens the influence of the number of actives stores over the average sales of the stores, and (2) the effect of seasonality is less important due to the absence of jackpot accumulations.

The next figure shows the distribution of the residuals for each of the layers. As can be seen, in all layers the residuals have a mean of zero, with distributions more concentrated around zero than in the case of normal distributions:

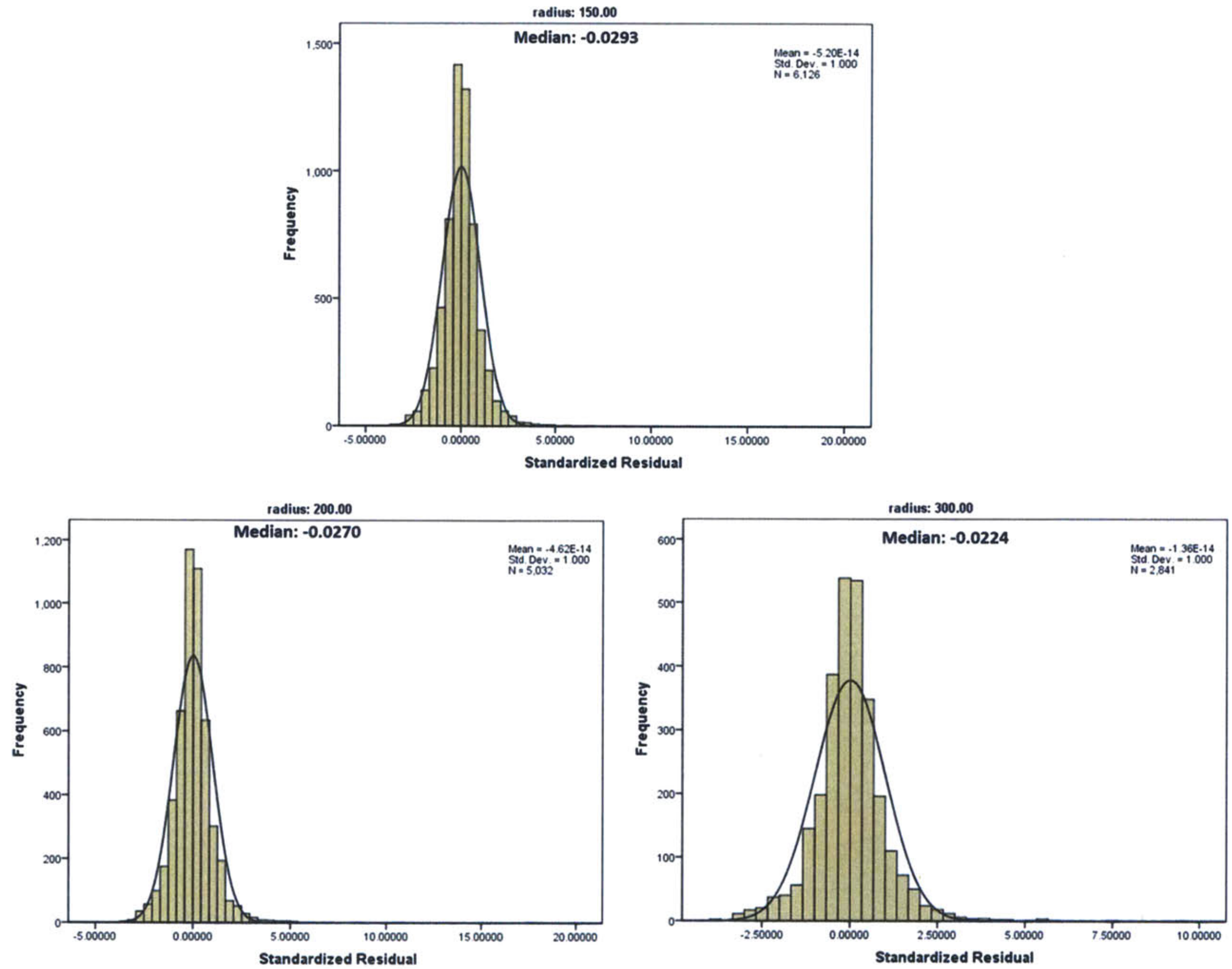

Figure 28: Standardized Residuals, Instant Games, Linear Model

Regressions were also run using the predicted values as dependent variables and the residuals as independent. In every case, these regressions showed non-significant results (both R^2 and F-test values were zero in every case), showing that errors are independent, confirming the significance of the original regressions.

The following table summarizes the coefficients obtained, along with the results of the bootstrap:

Table 19: Summary of Results with Coefficients, Aggregated Analysis, Instant Games, Linear Model

		Constant			Active Normalized (cannibalization coefficient)			Seasonal Factor		
	Radius	**150**	**200**	**300**	**150**	**200**	**300**	**150**	**200**	**300**
Regression	Unstandardized Coefficients	0.595	0.552	0.500	-0.476	-0.426	-0.374	0.864	0.857	0.860
	Std. Error	0.045	0.042	0.042	0.044	0.041	0.040	0.018	0.019	0.024
	Standardized Coefficients				-0.120	-0.128	-0.150	0.528	0.538	0.573
	t	13.174	13.084	11.866	-10.856	-10.470	-9.307	47.855	44.135	35.523
	Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bootstrap	Bias	0.000	0.002	0.001	0.000	-0.001	-0.002	0.000	-0.001	0.001
	Std. Error	0.044	0.051	0.048	0.043	0.048	0.052	0.018	0.020	0.024
	Sig. (2-tailed)	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	95% interval, lower	0.511	0.452	0.410	-0.562	-0.523	-0.474	0.828	0.818	0.815
	95% interval, Upper	0.676	0.647	0.592	-0.387	-0.334	-0.269	0.896	0.898	0.911

As seen in the table, the coefficients for Active Normalized and Seasonal Factor were significant, and the bootstrap shows that their estimation seems to be robust, as the 95% confidence intervals are relatively small around the values estimated by least squares. From this analysis, the following conclusions can be drawn:

- The signs of the coefficients are aligned with what was expected:
  - Positive coefficients for the seasonal factors, which indicate that stores increase their sales when the season is high.
  - Negative coefficients for the number of active nearby stores (Active Normalized), which indicates the presence of cannibalization, although the magnitude of the coefficients and values for the Adjusted R^2's, both smaller than in the case of lottery games, suggest that cannibalization is less important in this product category. As was mentioned before, this was an expected result and aligned with the logic of the phenomena: the purchase process of instant games is mostly driven by impulse, which means than an increase in the number of stores increases the likelihood of customers running into "opportunities" where impulse can drive a purchase. This is different from what happens with lottery games, where the purchase is usually planned, making it, therefore, more prone to cannibalization.
- Although relatively small, there seems to be a correlation again between the size of the cannibalization coefficients and the radius. As the radius increases, coefficients tend to become less negative, which in this case is aligned with logic (the farther the stores, the less likely that they will "steal" each other's customers for purchases driven by impulse). The bootstrap intervals, however, overlap, suggesting that these apparent differences may not be significant, and may be the results of just more observations within each geocluster with changes in the

number of active stores, although in this case, the parameter estimated by means squares and the intervals' lower and upper limits show gradients in the expected direction.

In the case of the differences model, the following results were obtained:

Table 20: Summary of Results, Aggregated Analysis, Instant Games, Differences Model

Radius [m]	R	R^2	Adjusted R^2	Std. Error of the Estimate	F test	F test prob.
150	0.558	0.312	0.311	0.392	1253.483	0
200	0.570	0.325	0.325	0.376	1097.583	0
300	0.583	0.340	0.339	0.353	664.129	0

As seen, the differences model delivered slightly better results for every case, although the Adjusted R^2's were still lower. The regression, however, produced statistically significant results, as evidenced by the values of the F-tests.

The next figure shows the distribution of the residuals for each of the layers. As seen, in all layers the residuals have a mean of zero, with distributions more concentrated around zero than in the case of normal distributions:

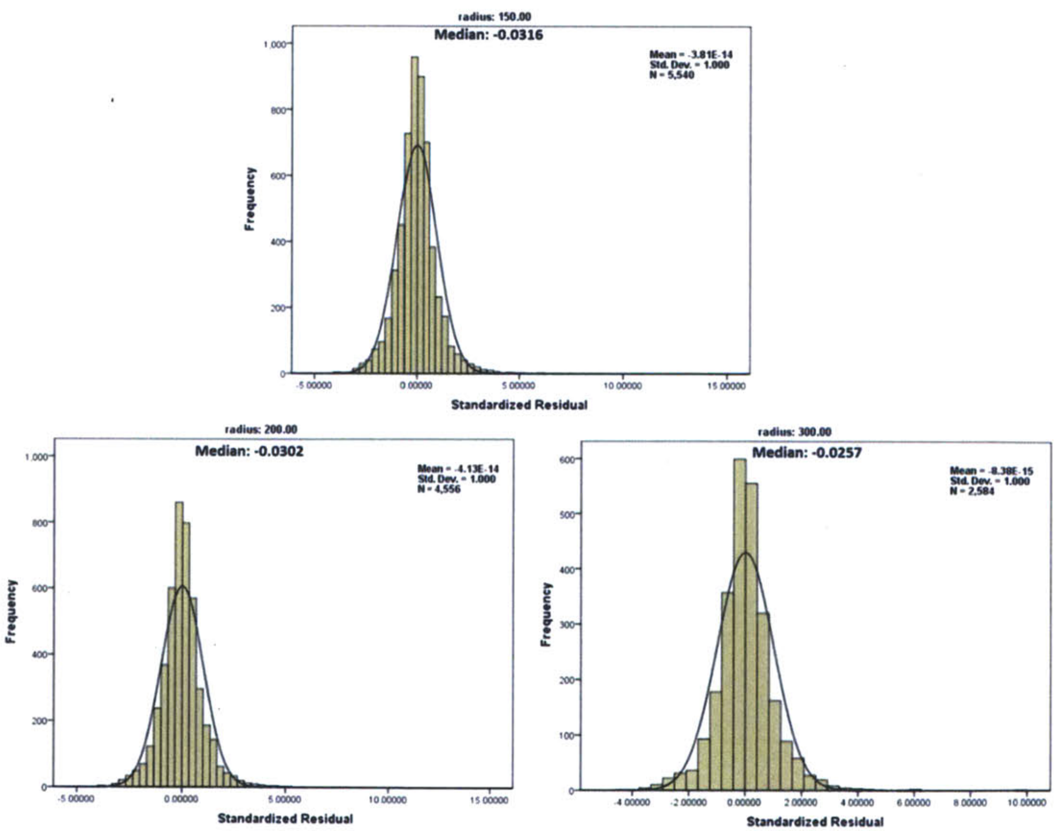

Figure 29: Standardized Residuals, Lottery Games, Instant Model

Regressions also were run using the predicted values as dependent variables and the residuals as independent. In every case, they showed non-significant results (both R^2 and F-test values were zero in every case), confirming the significance of the regressions.

The following table summarizes the coefficients obtained, along with the results of the bootstrap:

Table 21: Summary of Results with Coefficients, Aggregated Analysis, Instant Games, Differences Model

		Constant			Active Normalized (cannibalization coefficient)			Seasonal Factor		
	Radius	**150**	**200**	**300**	**150**	**200**	**300**	**150**	**200**	**300**
Regression	Unstandardized Coefficients	0.619	0.571	0.483	-0.510	-0.460	-0.359	0.880	0.878	0.866
Regression	Std. Error	0.045	0.041	0.042	0.043	0.040	0.040	0.018	0.019	0.024
Regression	Standardized Coefficients				-0.134	-0.143	-0.147	0.563	0.579	0.601
Regression	t	13.915	13.759	11.372	-11.853	-11.571	-8.941	49.887	46.744	36.445
Regression	Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bootstrap	Bias	-0.003	0.000	0.004	0.002	0.000	-0.005	0.001	0.000	0.001
Bootstrap	Std. Error	0.045	0.043	0.048	0.042	0.042	0.050	0.018	0.018	0.024
Bootstrap	Sig. (2-tailed)	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Bootstrap	95% interval, lower	0.527	0.480	0.399	-0.594	-0.544	-0.463	0.845	0.843	0.824
Bootstrap	95% interval, Upper	0.704	0.653	0.584	-0.428	-0.374	-0.261	0.916	0.915	0.917

As seen in the table, coefficients for Active Normalized and Seasonal Factor were again significant, and the bootstrap shows that their estimation seems to be robust, as the 95% confidence intervals are relatively small around the values estimated by the least squares. From this analysis, the following conclusions can be drawn:

- The signs of the coefficients are aligned with what was expected:
    - Positive coefficients for the seasonal factors, which indicate that stores increase their sales when the season is high.
    - Negative coefficients for the number of active nearby stores (Active Normalized), which again indicates the presence of cannibalization, although the magnitude of the coefficients and values for the Adjusted R^2's, both smaller than in the case of lottery games, suggest that cannibalization is less important in this product category. As mentioned before, this was an expected result and aligned with the logic of the situation: the purchase process in the case of instant games is mostly driven by impulse, which means that an increase in the number of stores increases the likelihood of customers running into "opportunities" where impulse can drive a purchase. This is different from what happens with lottery games where purchases are usually planned, making it, therefore, more prone to cannibalization.
- Although relatively small, there seems to be a correlation between the size of the cannibalization coefficient and the radius. As the radius increases, the coefficients tend to become less negative, which in this case is aligned to what was expected (the farther the stores,

the less likely that they will "steal" each other's customers for purchases driven by impulse). The bootstrap intervals, however, overlap, suggesting that these apparent differences may not be significant, and may be the results of just more observations within each geocluster with changes in the number of active stores, although in this case, the parameter estimated by means squares and the intervals' lower and upper limit show gradients in the expected direction.

The following results were obtained using exponential models. They are shown, as in the case of lottery games, just for illustrative purposes because their coefficients can be interpreted as elasticities:[11]

Table 22: Cannibalization Elasticities, Instant Games

	Constant			Active Normalized (cannibalization coefficient)			Seasonal Factor		
Radius	**150**	**200**	**300**	**150**	**200**	**300**	**150**	**200**	**300**
Unstandardized Coefficients	-0.141	-0.131	-0.115	-0.440	-0.346	-0.247	1.060	1.047	1.019
Std. Error	0.008	0.009	0.011	0.067	0.061	0.060	0.024	0.026	0.033
Standardized Coefficients				-0.074	-0.071	-0.069	0.496	0.505	0.516
t	-17.535	-15.254	-10.642	-6.584	-5.705	-4.115	44.076	40.698	31.020
Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

As mentioned, coefficients in the exponential model can be directly interpreted as elasticities: A coefficient of -0.44 for the case of the 150 m radius, for example, indicates that an increase of 1% in the concentration of active stores is expected to cause a decrease of 0.44% in the average size of the stores. In this regard, these coefficients provide a more intuitive interpretation.

A subsequent section in this chapter will provide results at the level of the demand clusters calculated in Chapter 2.

[11] Adjusted R^2's for the 150, 200 and 300 m radius were 0.242, 0.248 and 0.255 respectively.

Aggregated Comparison between Lottery and Instant Games

As seen in the previous sections, models for lottery games consistently delivered more significant results than the ones for instant games. This was a foreseeable situation, as the two independent variables used in the models, Active Normalized (measuring the number of active neighbors in each geocluster) and Seasonal Factor (capturing seasonal effects), were expected to have less influence over the dependent variable, the average sales per store, for the case of instant games. Purchase decisions for instant games are mostly driven by impulse, and seasonality is less relevant (as there are no jackpot accumulations), lowering the models' precision. Regressions, however, delivered in both cases significant results, as the t-tests for the coefficients were significant, and the bootstrap gave intervals for the parameters that had a relatively small range and did not contain zeros. The signs of the coefficients were also, in both product categories and for every layer (150, 200 and 300 m), aligned with logic, confirming the presence of cannibalization (as the coefficients for Active Normalized were always negative) and the influence of seasonality.

The intent of this section is to provide an easier comparison of the results for both product categories, through visual representations that include them both. In this regard, the following graph shows, for the case of the linear model, a comparison of the coefficients measuring the cannibalization effect (Active Normalized) and their lower and upper limits for the 95% interval obtained using bootstrapping:

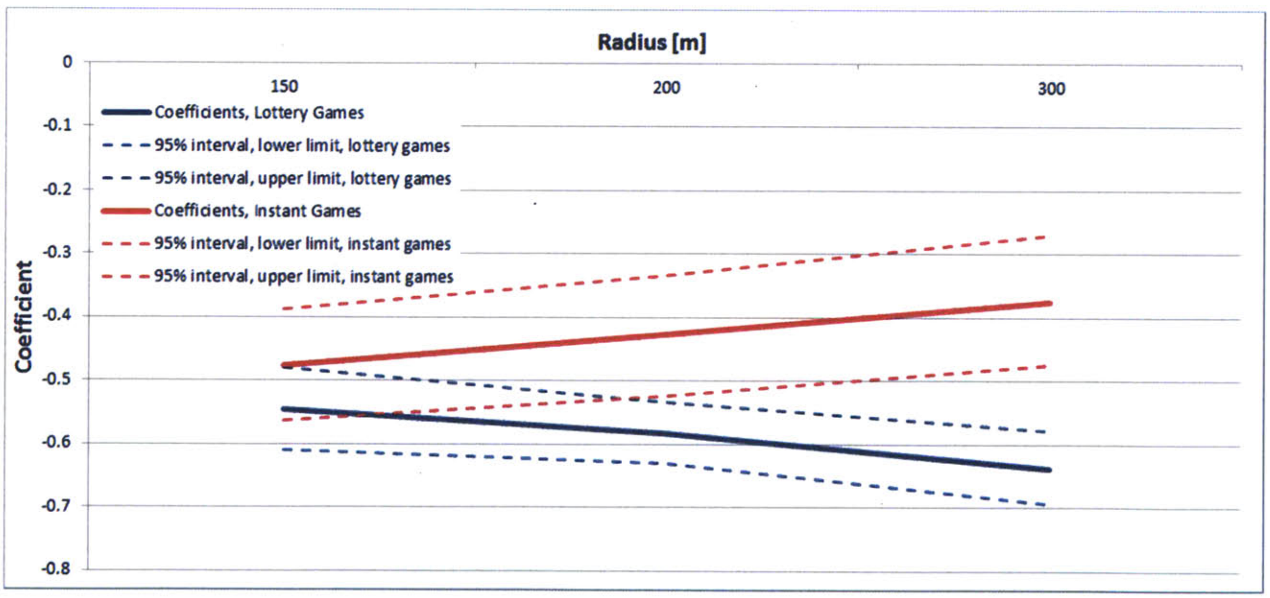

Figure 30: Comparison of Cannibalization Coefficients, Linear Models

The following conclusions can be drawn from inspecting this graph:

- Coefficients for instant games are consistently lower (in absolute terms) than the ones obtained for lottery games, denoting that cannibalization is lower for this product category, as its purchase process is mostly driven by impulse.
- The ranges of the intervals obtained for instant games are consistently bigger (both in absolute and percentage terms). This is aligned with what was observed for the models' fit

measurements in both cases: Adjusted R^2's were lower for instant games and, therefore, more unstable parameters were expected.

- In both cases, there are gradients associated with the radiuses, but they trend in different directions: for instant games, cannibalization coefficients decrease in severity as the radius increases, while for lottery games cannibalization seems to increase. For instant games, the situation makes sense: for an impulse purchase product category, the farther away the stores are, the lower should the cannibalizations be. For lottery games, the relationship is less intuitive, and the hypothesis is that the increase (in absolute terms) of the magnitude of the coefficients is explained more by the presence of additional observations within the geoclusters including variations in the number of active stores, than by a real increase in cannibalization. In any case, percentage-wise, variations are far more important for instant games than for lottery games: in the first case, the cannibalization coefficient decreases (absolute value) 21% from 150 to 300 m, while in the second case, the increase (absolute value) is 16%.

The following graph shows a similar comparison, but for the coefficients obtained using the differences model:

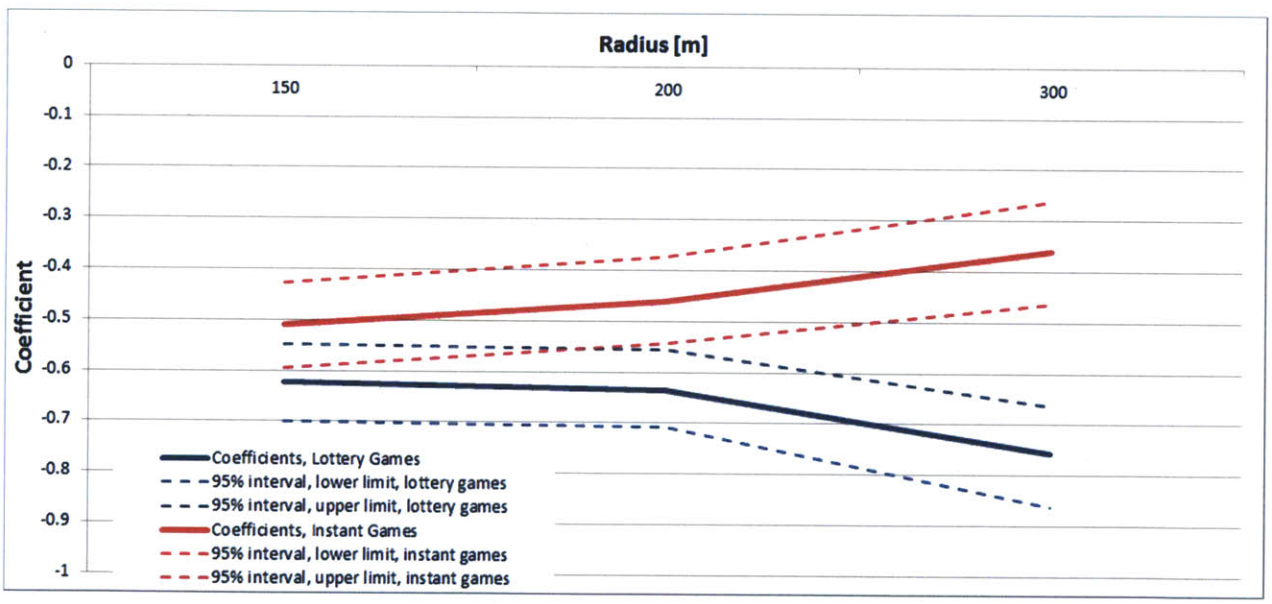

Figure 31: Comparison of Cannibalization Coefficients, Differences Models

The same conclusions can be drawn from this graph than in the previous case, although there are two aspects that are worth noticing:

- The ranges of the intervals are smaller, which is aligned with the consistently higher values obtained for the Adjusted R^2's.
- The slopes of the curves are slightly steeper: the coefficient for instant games decreases (in absolute value) 30% from 150 to 300 m and the increase (also in absolute value) in the coefficient for lottery games is now 21% (also from 150 to 300 m).

In more general terms, both product categories exhibit signs of saturation in their distribution network, as there is sufficient evidence suggesting that increases in the number of stores have caused a diminution in the average sales of each store. The analysis shows this relationship clearly and cleanly from seasonal effects. In this regard, the company's sales are not met with an infinite demand or, in practical terms, a demand that is much higher than current sales, which make the location of future stores a sensitive decision, as they may end up being a redistribution game (near-to-zero sum game), in which sales are just reallocated, as opposed to a real contribution to increase overall sales. Furthermore, the cannibalization measurement presented in this chapter can be thought of as a number reflecting the state of the store network in terms of saturation, allowing comparisons across product categories, industries, companies and territories.

The following section replicates this analysis at the level of demand clusters. The results of the disaggregated analysis, however, although more conceptually precise (as demand potential differences were controlled using the clustering procedure described in Chapter 2), had the problem that the number of observations with differences in the number of actives stores in some cases was too low, hindering the stability and significance of the parameters estimated.

Modeling Cannibalization at the Demand Cluster Level

The next step was to measure cannibalization at the level of demand clusters. As mentioned at the end of the previous section, a disaggregated analysis has the advantage of controlling by demand potential. The disadvantage, however, is that fewer cases were available for the estimation process, which hindered the statistical significance of the calculation of some of the parameters, particularly in demand clusters with few observations and/or where there was less history of variations in the number of active stores. Trying to avoid this situation, clusters 3 and 5 were fused (cluster 3, "medium socioeconomic level, high concentration of transportation and service points" and cluster 5, "medium socioeconomic level, high concentration of transportation points and residential"). Both clusters originally had similar demand characteristics (were "close clusters"), and cluster 3 had the disadvantage of having very few variations in the number of active stores in its geoclusters. Because of this, a new cluster was created with the cases from cluster 3 and 5; it was labeled as "medium socioeconomic level, high concentration of transportation points and residential," and was assigned the id 35.

The same models were used in this case, but the parameters' estimation was made inside each of the clusters:

- A linear model: $t_{ij} = \alpha \cdot a_{ij} + \beta \cdot f_i + C$
- A differences model: $\Delta t_{ij} = \alpha \cdot \Delta a_{ij} + \beta \cdot \Delta f_i + C$ were $\Delta X = X_{t-1} - X_t$

Accordingly, this section reports on the results obtained for each of the combinations of demand clusters, product category and geocluster layers (150, 200 and 300 m). Results were grouped in subsections, based on the product category and radius layers.

Lottery Games

150 m Layer

The following results were obtained for the estimation of the linear models, for the 150 m layer:

Table 23: Summary of Results, Disaggregated Analysis, Lottery Games, 150 m, Linear Model

Id	Cluster Description	R	R^2	Adj. R^2	Std. Error of the Estimate	F	Sig.
1	High socioeconomic level, high concentration of commerce, transportation and service points, low population density	0.871	0.758	0.758	0.29112	1499	0
2	Medium-high socioeconomic level, with concentration of commerce, transportation and service points	0.895	0.801	0.8	0.23073	2488	0
4	Medium socioeconomic level, high concentration of transportation points and highly residential	0.872	0.761	0.759	0.28961	485	0
35	Medium socioeconomic level, high concentration of transportation points and residential	0.923	0.851	0.851	0.19654	1207	0
6	Medium socioeconomic level, somewhat residential with few points of interest	0.904	0.818	0.817	0.20638	2019	0
7	Medium socioeconomic level, not very residential with very low population density	0.895	0.801	0.8	0.23808	976	0
8	Medium socioeconomic level, highly residential with very few points of interest	0.907	0.822	0.821	0.20586	921	0
9	Medium socioeconomic level, very high concentration of commerce, transportation and service points	0.892	0.795	0.792	0.21345	295	0
10	Medium-low socioeconomic level, pure residential	0.901	0.812	0.811	0.20641	1145	0
11	Medium-low socioeconomic level, pure residential with low density	0.87	0.756	0.753	0.24016	242	0
12	Medium-low socioeconomic level, with high concentration of commerce, transportation and service points, low density	0.88	0.775	0.775	0.23477	1972	0
13	Low socioeconomic level, residential with high population density	0.865	0.749	0.748	0.28092	847	0
14	Low socioeconomic level, residential with high population density, some concentration of transportation points	0.859	0.737	0.736	0.2974	801	0
15	Low socioeconomic level, residential with medium population density	0.884	0.782	0.781	0.2208	745	0

As seen in the table, the linear model delivered Adjusted R^2's in the range [0.74, 0.85] for all the clusters, and in every case the F-tests indicated that regressions were significant.

The following table shows the results obtained for the parameters, including the intervals delivered by a bootstrap procedure with 500 samples. To simplify the table, only the value of the parameter is listed for the constant.

Table 24: Summary of Results with Coefficients, Disaggregated Analysis, Lottery Games, 150 m, Linear Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.306	0.036	-0.584	0.013	1.088	0.002	1.004	1.186	0.205	*
2	-0.639	0.002	-0.733	-0.556	1.036	0.002	0.98	1.091	0.592	
4	-0.965	0.002	-1.398	-0.59	0.858	0.002	0.743	0.958	1.064	
35	-0.82	0.002	-0.999	-0.664	0.944	0.002	0.884	1.015	0.86	
6	-0.514	0.002	-0.701	-0.268	1.014	0.002	0.954	1.071	0.49	
7	-0.576	0.002	-0.722	-0.442	1.001	0.002	0.915	1.111	0.561	
8	-0.718	0.002	-0.941	-0.438	0.941	0.002	0.857	1.013	0.758	
9	-0.674	0.002	-0.906	-0.51	1.204	0.002	1.009	1.421	0.467	
10	-0.767	0.002	-0.985	-0.584	1.029	0.002	0.952	1.11	0.732	
11	-0.361	0.218	-0.934	0.24	1.02	0.002	0.91	1.132	0.33	*
12	-0.461	0.002	-0.557	-0.359	1.02	0.002	0.963	1.078	0.431	
13	-0.579	0.002	-0.722	-0.412	0.889	0.002	0.817	0.969	0.657	
14	-0.427	0.034	-0.799	0.028	0.882	0.002	0.822	0.945	0.496	*
15	-0.3	0.126	-0.603	0.361	1.168	0.002	1.071	1.27	0.124	*

Clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This occurs most likely due to few geoclusters within the demand cluster with variations in the number of active stores. This was the case of clusters 1, 11, 14 and 15.

The following results were obtained for the differences model, also for the 150 m layer (the name of the clusters will be subsequently omitted, for simplicity):

Table 25: Summary of Results, Disaggregated Analysis, Lottery Games, 150 m, Differences Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.884	0.782	0.781	0.281	1586.277	0
2	0.917	0.841	0.841	0.208	3047.247	0
4	0.924	0.854	0.853	0.224	814.365	0
35	0.939	0.882	0.882	0.176	2430.974	0
6	0.924	0.854	0.853	0.186	1003.500	0
7	0.904	0.818	0.817	0.231	1067.959	0
8	0.924	0.853	0.852	0.190	465.434	0
9	0.932	0.868	0.867	0.173	1232.116	0
10	0.913	0.834	0.834	0.195	446.343	0
11	0.928	0.861	0.859	0.184	2141.518	0
12	0.895	0.802	0.801	0.222	1147.943	0
13	0.903	0.815	0.815	0.242	988.119	0
14	0.890	0.793	0.792	0.265	896.857	0
15	0.908	0.824	0.823	0.199	1465.745	0

As seen in the table, the differences model delivered Adjusted R^2's in the range [0.78, 0.88] for all the clusters, and in every case, the F-tests indicated that regressions were significant. Adjuster R^2's were in this case slightly higher than in the case of the linear models. The following table shows the results obtained for the parameters:

Table 26: Summary of Results with Coefficients, Disaggregated Analysis, Lottery Games, 150 m, Differences Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.310	0.000	-0.603	0.074	1.092	0.000	1.004	1.183	0.213	*
2	-0.661	0.000	-0.747	-0.581	1.053	0.000	1.007	1.099	0.608	
4	-0.940	0.000	-1.435	-0.526	0.902	0.000	0.824	0.988	1.021	
35	-0.828	0.000	-1.019	-0.670	0.955	0.000	0.899	1.020	0.868	
6	-0.608	0.000	-0.789	-0.451	1.015	0.000	0.960	1.067	0.593	
7	-0.571	0.000	-0.708	-0.442	0.996	0.000	0.898	1.108	0.570	
8	-0.749	0.000	-1.028	-0.463	0.947	0.000	0.860	1.030	0.793	
9	-0.715	0.000	-0.980	-0.541	1.288	0.000	1.124	1.487	0.440	
10	-0.781	0.000	-1.004	-0.592	1.037	0.000	0.962	1.119	0.748	
11	-0.870	0.000	-1.042	-0.737	1.076	0.000	0.980	1.198	0.803	
12	-0.482	0.000	-0.581	-0.369	1.025	0.000	0.960	1.091	0.454	
13	-0.659	0.000	-0.817	-0.511	0.907	0.000	0.837	0.996	0.739	
14	-0.510	0.000	-0.852	-0.045	0.889	0.000	0.820	0.958	0.599	*
15	-0.304	0.097	-0.625	0.413	1.176	0.000	1.084	1.280	0.132	*

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 1, 14 and 15. Cluster 11 with this model delivered a significant parameter (unlike what happened for the linear model).

200 m Layer

The following results were obtained for the estimation with linear models, for the 200 m layer:

Table 27: Summary of Results, Disaggregated Analysis, Lottery Games, 200 m, Linear Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	.870	0.758	0.757	0.28604	1289.709	0
2	.893	0.798	0.798	0.22745	1846.464	0
4	.872	0.761	0.759	0.2921	427.538	0
35	.928	0.86	0.86	0.1957	1149.54	0
6	.902	0.813	0.812	0.20886	1487.176	0
7	.897	0.804	0.803	0.2459	789.424	0
8	.908	0.825	0.824	0.21049	848.837	0
9	.921	0.849	0.845	0.14292	199.665	0
10	.919	0.844	0.843	0.18351	1288.151	0
11	.862	0.744	0.738	0.25155	134.897	0
12	.898	0.807	0.806	0.20045	1897.616	0
13	.870	0.757	0.756	0.27434	783.247	0
14	.883	0.779	0.779	0.25957	960.908	0
15	.900	0.81	0.809	0.19827	725.023	0

As seen in the table, the linear model delivered Adjusted R^2's in the range [0.74, 0.86] for all the clusters, and in every case the F-tests indicated that regressions were significant.

The following table shows the results obtained for the parameters, including the intervals delivered by a bootstrap procedure with 500 samples:

Table 28: Summary of Results with Coefficients, Disaggregated Analysis, Lottery Games, 200 m, Linear Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	Mark
1	-0.378	0.000	-0.590	-0.082	1.076	0.000	0.982	1.172	0.292	*
2	-0.658	0.000	-0.757	-0.570	1.107	0.000	1.043	1.168	0.538	
4	-0.834	0.000	-1.029	-0.657	0.810	0.000	0.675	0.938	0.981	
35	-0.675	0.000	-0.802	-0.555	0.952	0.000	0.889	1.016	0.710	
6	-0.464	0.000	-0.768	-0.107	1.081	0.000	1.002	1.159	0.374	
7	-0.673	0.000	-0.827	-0.545	1.041	0.000	0.929	1.166	0.616	
8	-0.610	0.000	-0.849	-0.296	0.960	0.000	0.871	1.049	0.628	
9	-0.573	0.000	-0.761	-0.423	0.971	0.000	0.850	1.056	0.613	
10	-0.724	0.000	-1.006	-0.536	1.006	0.000	0.943	1.076	0.714	
11	-0.449	0.008	-1.104	0.238	0.753	0.000	0.652	0.852	0.649	*
12	-0.628	0.000	-0.722	-0.522	0.984	0.000	0.925	1.047	0.642	
13	-0.566	0.000	-0.717	-0.432	0.962	0.000	0.877	1.073	0.575	
14	-0.523	0.000	-0.663	-0.384	0.884	0.000	0.827	0.946	0.600	
15	-0.270	0.001	-0.499	-0.011	1.098	0.000	1.008	1.199	0.170	*

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This occurs most likely due to few geoclusters within the demand cluster with variations in the number of active stores. This was the case of clusters 1, 11 and 15. The estimations for 200 m showed, as expected, more robust results than the one for 150 m.

The following results were obtained for the differences model, also for the 200 m layer:

Table 29: Summary of Results, Disaggregated Analysis, Lottery Games, 200 m, Differences Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.88	0.774	0.773	0.2812	1312.241	0
2	0.92	0.847	0.847	0.1999	2388.917	0
4	0.928	0.86	0.859	0.2216	758.211	0
35	0.941	0.886	0.885	0.1790	1734.93	0
6	0.92	0.846	0.845	0.1917	842.049	0
7	0.908	0.825	0.824	0.2349	1027.748	0
8	0.928	0.861	0.86	0.1904	201.063	0
9	0.927	0.859	0.855	0.1420	1320.884	0
10	0.926	0.857	0.856	0.1790	330.693	0
11	0.941	0.886	0.883	0.1654	1999.897	0
12	0.908	0.825	0.825	0.1933	1051.258	0
13	0.905	0.82	0.819	0.2377	1208.695	0
14	0.911	0.83	0.829	0.2276	884.293	0
15	0.921	0.849	0.848	0.1786	1341.565	0

As seen in the table, the differences model delivered Adjusted R^{2}'s in the range [0.78, 0.89] for all the clusters, and in every case, the F-tests indicated that the regressions were significant. Adjusted R^{2}'s were slightly higher than in the case of the linear models, and higher also than in the estimations for 150m. The following table shows the results obtained for the parameters:

Table 30: Summary of Results with Coefficients, Disaggregated Analysis, Lottery Games, 200 m, Differences Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.386	0.000	-0.628	-0.089	1.079	0.000	0.985	1.183	0.303	*
2	-0.702	0.000	-0.803	-0.609	1.129	0.000	1.072	1.193	0.575	
4	-0.832	0.000	-1.017	-0.660	0.859	0.000	0.782	0.959	0.958	
35	-0.684	0.000	-0.825	-0.557	0.966	0.000	0.899	1.044	0.715	
6	-0.612	0.000	-0.893	-0.350	1.081	0.000	1.012	1.157	0.530	
7	-0.683	0.000	-0.832	-0.574	1.039	0.000	0.932	1.165	0.639	
8	-0.644	0.000	-0.928	-0.282	0.967	0.000	0.879	1.053	0.664	
9	-0.579	0.000	-0.760	-0.431	0.972	0.000	0.885	1.079	0.617	
10	-0.736	0.000	-0.987	-0.534	1.012	0.000	0.947	1.083	0.727	
11	-0.924	0.000	-1.287	-0.610	0.782	0.000	0.711	0.872	1.127	
12	-0.649	0.000	-0.753	-0.550	0.989	0.000	0.931	1.056	0.662	
13	-0.609	0.000	-0.726	-0.494	0.985	0.000	0.900	1.099	0.614	
14	-0.585	0.000	-0.708	-0.465	0.879	0.000	0.820	0.940	0.690	
15	-0.289	0.000	-0.529	-0.018	1.105	0.000	1.020	1.202	0.190	*

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 1 and 15. Cluster 11 with this model delivered a significant parameter (unlike what happened for the linear model). The estimation for cluster 14 was in this case also significant (in the 150 m it was not).

300 m Layer

The following results were obtained for the estimation with linear models, for the 300 m layer:

Table 31: Summary of Results, Disaggregated Analysis, Lottery Games, 300 m, Linear Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.901	0.811	0.811	0.2561	998.605	0
2	0.904	0.817	0.816	0.2158	976.271	0
4	0.869	0.754	0.752	0.2516	310.403	0
35	0.933	0.871	0.87	0.1708	856.696	0
6	0.902	0.814	0.813	0.2032	774.226	0
7	0.926	0.857	0.855	0.1742	625.066	0
8	0.92	0.846	0.844	0.2139	525.533	0
9	0.967	0.935	0.931	0.0911	210.131	0
10	0.946	0.894	0.893	0.1569	946.116	0
11	0.91	0.828	0.814	0.2227	62.413	0
12	0.916	0.838	0.838	0.1837	1406.365	0
13	0.848	0.72	0.717	0.3254	331.158	0
14	0.864	0.747	0.745	0.2694	483.654	0
15	0.91	0.828	0.826	0.1732	586.006	0

As seen in the table, the linear model delivered Adjusted R^2's in the range [0.72, 0.93] for all the clusters, the highest average range of any layer with linear models, and in every case the F-tests indicated that the regressions were significant.

The following table shows the results obtained for the parameters, including the intervals delivered by a bootstrap procedure with 500 samples:

Table 32: Summary of Results with Coefficients, Disaggregated Analysis, Lottery Games, 300 m, Linear Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	**B**	**Sig.**	95% BT LL	95% BT UP	**B**	**Sig.**	95% BT LL	95% BT LL	**B**	
1	-0.746	0.000	-0.919	-0.552	1.102	0.000	0.998	1.213	0.631	
2	-0.728	0.000	-0.936	-0.563	1.112	0.000	1.018	1.212	0.601	
4	-0.626	0.000	-0.806	-0.508	0.781	0.000	0.608	0.917	0.826	
35	-0.674	0.000	-0.835	-0.545	0.915	0.000	0.846	1.003	0.750	
6	-0.446	0.000	-0.628	-0.249	1.019	0.000	0.919	1.113	0.422	
7	-0.827	0.000	-0.965	-0.659	0.973	0.000	0.898	1.060	0.841	
8	-0.761	0.000	-0.978	-0.576	0.972	0.000	0.874	1.084	0.758	
9	-0.727	0.000	-1.113	-0.549	0.826	0.000	0.690	1.134	0.902	
10	-0.824	0.000	-1.294	-0.505	1.042	0.000	0.958	1.124	0.778	
11	-1.184	0.001	-1.999	-0.841	0.690	0.000	0.586	0.919	1.390	
12	-0.687	0.000	-0.825	-0.569	0.983	0.000	0.905	1.065	0.705	
13	-0.671	0.000	-0.870	-0.454	0.996	0.000	0.853	1.143	0.624	
14	-0.382	0.000	-0.533	-0.270	0.920	0.000	0.820	1.032	0.439	
15	-0.579	0.000	-0.767	-0.360	1.023	0.000	0.904	1.134	0.555	

In this case, all the cannibalization coefficients were significant and relatively stable, even for clusters 1 and 15, the most problematic in all other estimations. This is because, as explained before, this layer is the one that has the most evidence in the history it covers of geoclusters with variations in the number of active stores.

The following results were obtained for the differences model, also for the 300 m layer:

Table 33: Summary of Results, Disaggregated Analysis, Lottery Games, 300 m, Differences Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.91	0.829	0.829	0.2483	1047.385	0
2	0.93	0.855	0.855	0.1936	1197.339	0
4	0.94	0.884	0.883	0.1731	714.068	0
35	0.95	0.907	0.906	0.1465	1145.112	0
6	0.91	0.832	0.831	0.1971	816.542	0
7	0.95	0.902	0.901	0.1449	890.39	0
8	0.95	0.894	0.893	0.1802	746.048	0
9	0.97	0.935	0.93	0.0936	193.587	0
10	0.95	0.901	0.9	0.1546	955.507	0
11	0.94	0.886	0.876	0.1731	89.695	0
12	0.92	0.846	0.845	0.1831	1394.702	0
13	0.89	0.789	0.787	0.2856	441.8	0
14	0.89	0.79	0.789	0.2449	564.072	0
15	0.92	0.854	0.853	0.1619	664.571	0

As seen in the table, the differences model delivered Adjusted R^2's in the range [0.79, 0.93], for all the clusters, and in every case the F-tests indicated that regressions were significant. Adjusted R^2's were slightly higher than in the case of the linear models, and higher also than in the estimations for 150 m and 200 m. The following table shows the results obtained for the parameters:

Table 34: Summary of Results with Coefficients, Disaggregated Analysis, Lottery Games, 300 m, Differences Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	Mark
1	-0.767	0.000	-0.947	-0.575	1.104	0.000	0.997	1.224	0.658	
2	-0.749	0.000	-0.941	-0.565	1.128	0.000	1.035	1.242	0.620	
4	-0.645	0.000	-0.806	-0.515	0.847	0.000	0.783	0.927	0.795	
35	-0.696	0.000	-0.884	-0.547	0.939	0.000	0.879	1.018	0.757	
6	-0.480	0.000	-0.655	-0.293	1.023	0.000	0.930	1.131	0.453	
7	-0.805	0.000	-0.964	-0.644	0.974	0.000	0.906	1.055	0.826	
8	-0.803	0.000	-1.032	-0.640	0.989	0.000	0.899	1.108	0.799	
9	-0.729	0.000	-1.161	-0.551	0.821	0.000	0.704	1.168	0.911	
10	-0.825	0.000	-1.363	-0.488	1.038	0.000	0.965	1.127	0.785	
11	-1.117	0.000	-1.685	-0.818	0.656	0.000	0.566	0.780	1.402	
12	-0.698	0.000	-0.816	-0.576	0.985	0.000	0.910	1.068	0.713	
13	-0.759	0.000	-0.951	-0.590	1.014	0.000	0.883	1.170	0.720	
14	-0.423	0.000	-0.572	-0.293	0.914	0.000	0.805	1.027	0.508	
15	-0.596	0.000	-0.792	-0.377	1.029	0.000	0.919	1.141	0.572	

In this case, again all the cannibalization coefficients were significant and relatively stable, even for clusters 1 and 15, the most problematic in all other estimations. This is because, as explained before, this layer is the one that has the most evidence in the history it covers of geoclusters with variations in the number of active stores.

Summary

As was expected, as the radius used for the geocluster grows, so does the significance of the statistical analysis. This situation evidences a trade-off of precisions: geographic precision (which increases with smaller radiuses) vs. statistical reliability (which increases with bigger radiuses). There is an additional element of trade-off: demand clusters. Aggregated estimations (presented at the beginning of the chapter) provided more observations (increasing the precision in the estimation of the coefficients), but disregard demand differences; disaggregated estimations (at the level of demand clusters), consider differences associated with demand potentials, but provide fewer observations within each cluster to estimate meaningful coefficients. To compare the results obtained in every case, expected error in the cannibalization measurement was defined as the range of the 95% interval delivered by the bootstraps, divided by 2 times the unstandardized parameter:

$$Expected\ Error = \pm\frac{range}{2 \cdot \beta}$$

The following table summarizes the expected errors estimated using the linear functions, for the demand clusters and the aggregated estimation:

Table 35: Summary of Errors, Linear Model, Lottery Games

	ID	150 m	200 m	300 m
Demand Clusters	1	98%	67%	25%
	2	14%	14%	26%
	4	42%	22%	24%
	35	20%	18%	22%
	6	42%	71%	42%
	7	24%	21%	19%
	8	35%	45%	26%
	9	29%	29%	39%
	10	26%	32%	48%
	11	163%	149%	49%
	12	21%	16%	19%
	13	27%	25%	31%
	14	97%	27%	34%
	15	161%	90%	35%
	Average	**57%**	**45%**	**31%**
Aggregated		12%	8%	9%

As expected, precision increases as the radius increases and the errors obtained at the aggregated level are almost 5 times lower than the average at the level of demand clusters. The average level, however, is highly influenced by the errors of the clusters where the coefficients were not statistically significant. If those clusters are removed from the calculation (cluster 1, 11, 14 and 15 for the 150 m layer and 1, 11 and 15 for the 200 m layer), the averages for the 150 and 200 m layers are reduced to 28% and 25% (the 300 m layer does not change).

The 200 m layer was selected for the location assessment presented in the next chapter for lottery games: it offers a good trade-off between statistical and geographic precision (the average error at the level of demand clusters is 25%[12]) and the aggregated error is only 8%. Once the 200 m layer was selected for next's chapter location assessment, it was necessary to define a criterion to choose the set of coefficients for each demand cluster, considering that in some of them the expected precision by using just the parameters estimated at the level of demand clusters was too low. The procedure used was as follows:

- If the expected error at the level of the demand cluster based on the cannibalization coefficients was equal, less or marginally superior to 25%, then the set of coefficients estimated for the demand cluster was used.
- If the expected error for a demand cluster based on the cannibalization coefficients was significantly superior to 25%, then the expected error of the "closest" demand cluster, if there was a cluster close enough in terms of demand characteristics, was examined. If it was less than 25%, then the coefficients for the closest demand cluster were used.
- In any other case, the coefficients of the aggregated estimation were used.

The following table presents a summary of this process:

[12] When clusters 1, 11 and 15 are not considered.

Table 36: Selection of the Coefficients per Cluster, Lottery Games

ID	Expected Error	Less than 25%?	Closest cluster	Error closest cluster	Closest Less than 25%?	Cluster of the final coefficients to use
1	67%	No	2	14%	Yes	2
2	14%	Yes				2
4	22%	Yes				4
35	18%	Yes				35
6	71%	No	7	21%	Yes	7
7	21%	Yes				7
8	45%	No	10	32%	No	Aggregated
9	29%	No	none close enough			Aggregated
10	32%	No	11	149%	No	Aggregated
11	149%	No	10	32%	No	Aggregated
12	16%	Yes				12
13	25%	Yes				13
14	27%	Yes				14
15	90%	No	14	27%	Yes	14

Linear models were preferred in this case because the linear functional forms obtained have properties that will make the location assessment presented in the next chapter easier to calculate. The differences model delivered results that in most cases were comparable to the ones obtained using the linear model; in cases where there was a difference, the increment in precision and Adjusted R^2's was marginal.

Instant Games

150 m Layer

The following results were obtained for the estimation with linear models, for the 150 m layer:

Table 37: Summary of Results, Disaggregated Analysis, Instant Games, 150 m, Linear Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.498	0.248	0.245	0.4377	93.922	0
2	0.526	0.276	0.275	0.3784	175.732	0
4	0.501	0.251	0.244	0.4017	34.536	0
35	0.616	0.38	0.376	0.3739	101.854	0
6	0.623	0.388	0.386	0.3597	214.952	0
7	0.592	0.351	0.347	0.3987	79.493	0
8	0.631	0.398	0.394	0.3064	106.074	0
9	0.624	0.389	0.379	0.3212	38.838	0
10	0.433	0.187	0.184	0.5670	49.459	0
11	0.594	0.353	0.341	0.3406	29.232	0
12	0.488	0.238	0.236	0.4487	133.688	0
13	0.486	0.237	0.233	0.4511	68.808	0
14	0.529	0.28	0.277	0.4111	91.752	0
15	0.656	0.43	0.427	0.3314	124.897	0

As seen in the table, the linear model delivered Adjusted R^2's in the range [0.19, 0.43], which are significantly lower than the ones obtained for lottery games, and also low in absolute terms. Removing the observation with the lowest value (cluster 10), however, the next lowest Adjusted R^2 is 0.233, and the average is 0.33. All the F-tests indicated that the regressions delivered significant results.

The following table shows the results obtained for the parameters, including the intervals delivered by a bootstrap procedure with 500 samples. To simplify the table, only the value of the parameter is listed for the case of the constant:

Table 38: Summary of Results with Coefficients, Disaggregated Analysis, Instant Games, 150 m, Linear Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	Mark
1	-0.735	0.000	-1.026	-0.501	0.783	0.000	0.690	0.894	0.935	
2	-0.559	0.000	-0.732	-0.356	0.889	0.000	0.800	0.974	0.663	
4	-1.190	0.002	-2.025	-0.505	0.629	0.000	0.458	0.790	1.523	
35	-0.745	0.000	-1.053	-0.480	0.977	0.000	0.835	1.122	0.751	
6	-0.353	0.003	-0.610	-0.093	0.944	0.000	0.877	1.028	0.391	*
7	0.022	0.886	-0.396	0.460	0.890	0.000	0.753	1.043	0.064	*
8	-0.586	0.000	-0.801	-0.343	0.839	0.000	0.728	0.938	0.731	
9	-0.861	0.001	-1.396	-0.383	0.888	0.000	0.647	1.146	0.968	
10	-0.726	0.011	-0.924	-0.556	0.894	0.000	0.744	1.089	0.807	
11	0.224	0.484	-0.509	0.850	0.676	0.000	0.479	0.892	0.082	*
12	-0.542	0.000	-0.723	-0.387	0.835	0.000	0.719	0.947	0.692	
13	-0.051	0.743	-0.316	0.240	0.545	0.000	0.463	0.628	0.466	*
14	-0.696	0.000	-1.136	-0.207	0.684	0.000	0.592	0.783	0.963	
15	-0.738	0.002	-1.488	-0.189	0.798	0.000	0.687	0.909	0.925	

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This occurs most likely due to few geoclusters within the demand cluster with variations in the number of active stores. This was the case of clusters 6, 7, 11 and 13. For this product category, impulse being the main purchase driver, cannibalization coefficients were expected also to be less significant.

The following results were obtained for the differences model, also for the 150 m layer:

Table 39: Summary of Results, Disaggregated Analysis, Instant Games, 150 m, Differences Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.525	0.275	0.273	0.4263	97.72	0
2	0.572	0.327	0.326	0.3545	204.168	0
4	0.529	0.28	0.272	0.3671	35.734	0
35	0.622	0.387	0.383	0.3718	95.833	0
6	0.64	0.41	0.408	0.3512	213.05	0
7	0.599	0.359	0.354	0.3999	74.354	0
8	0.673	0.453	0.45	0.2919	120.645	0
9	0.633	0.4	0.389	0.3244	37.03	0
10	0.551	0.304	0.3	0.4455	84.494	0
11	0.606	0.367	0.354	0.3445	27.841	0
12	0.493	0.243	0.241	0.4447	124.147	0
13	0.541	0.293	0.29	0.4188	82.706	0
14	0.53	0.281	0.277	0.4077	81.945	0
15	0.664	0.442	0.438	0.3301	118.584	0

As seen in the table, the differences model delivered Adjusted R^2's in the range [0.24, 0.45], but for all the clusters, the F-tests indicated that the regressions were significant. Adjusted R^2's were slightly higher than in the case of the linear models. The following table shows the results obtained for the parameters:

Table 40: Summary of Results with Coefficients, Disaggregated Analysis, Instant Games, 150 m, Differences Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.750	0.000	-1.044	-0.548	0.789	0.000	0.688	0.888	0.950	
2	-0.509	0.000	-0.699	-0.306	0.912	0.000	0.831	0.995	0.590	
4	-1.224	0.001	-2.231	-0.410	0.652	0.000	0.470	0.781	1.539	
35	-0.752	0.000	-1.077	-0.411	0.976	0.000	0.833	1.126	0.768	
6	-0.382	0.002	-0.652	-0.126	0.949	0.000	0.866	1.032	0.435	
7	-0.033	0.837	-0.483	0.451	0.901	0.000	0.747	1.063	0.122	*
8	-0.595	0.000	-0.842	-0.354	0.881	0.000	0.779	0.990	0.702	
9	-0.963	0.001	-1.411	-0.422	0.876	0.000	0.638	1.091	1.095	
10	-0.737	0.001	-0.994	-0.571	0.942	0.000	0.803	1.108	0.749	
11	-0.120	0.765	-0.695	0.322	0.738	0.000	0.508	0.946	0.366	*
12	-0.604	0.000	-0.768	-0.434	0.820	0.000	0.717	0.933	0.779	
13	-0.086	0.565	-0.331	0.255	0.583	0.000	0.498	0.664	0.469	*
14	-0.749	0.000	-1.132	-0.284	0.685	0.000	0.589	0.784	1.036	
15	-0.766	0.002	-1.494	-0.211	0.823	0.000	0.708	0.947	0.925	

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 7, 11 and 13. Cluster 6 with this model delivered a significant parameter (unlike what happened for the linear model).

200 m Layer

The following results were obtained for the estimation with linear models, for the 200 m layer:

Table 41: Summary of Results, Disaggregated Analysis, Instant Games, 200 m, Linear Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.493	0.243	0.24	0.4199	80.415	0
2	0.529	0.28	0.278	0.3671	139.497	0
4	0.556	0.309	0.302	0.3863	43.899	0
35	0.581	0.338	0.333	0.4081	71.681	0
6	0.646	0.418	0.415	0.3256	177.436	0
7	0.567	0.322	0.316	0.3773	54.029	0
8	0.63	0.396	0.392	0.3105	96.547	0
9	0.669	0.448	0.428	0.3079	22.287	0
10	0.432	0.187	0.183	0.5942	42.542	0
11	0.697	0.486	0.473	0.2894	38.257	0
12	0.546	0.298	0.296	0.3848	140.747	0
13	0.507	0.257	0.253	0.4215	68.676	0
14	0.502	0.252	0.248	0.4254	75.051	0
15	0.738	0.545	0.541	0.2862	159.153	0

As seen in the table, the linear model delivered Adjusted R^2's in the range [0.18, 0.43], which are significantly lower again than the ones obtained for lottery games, and also low in absolute terms. Removing the observation with the lowest value (cluster 10 again), however, the next lowest Adjusted R^2 is 0.24, and the average is 0.35. All the F-tests indicated that the regressions delivered significant results.

The following table shows the results obtained for the parameters, including the intervals delivered by a bootstrap procedure with 500 samples:

Table 42: Summary of Results with Coefficients, Disaggregated Analysis, Instant Games, 200 m, Linear Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.802	0.000	-1.127	-0.502	0.753	0.000	0.637	0.869	1.033	
2	-0.527	0.000	-0.814	-0.193	0.809	0.000	0.718	0.904	0.709	
4	-0.221	0.168	-0.673	0.335	0.653	0.000	0.523	0.786	0.516	*
35	-0.702	0.000	-1.054	-0.406	0.988	0.000	0.831	1.164	0.699	
6	-0.497	0.003	-0.906	-0.223	0.978	0.000	0.883	1.079	0.506	
7	-0.774	0.000	-1.034	-0.495	0.929	0.000	0.740	1.102	0.824	
8	-0.667	0.001	-1.089	-0.227	0.841	0.000	0.726	0.935	0.805	
9	-0.766	0.002	-1.108	-0.380	0.975	0.000	0.725	1.246	0.801	
10	-0.693	0.023	-0.902	-0.516	0.916	0.000	0.747	1.151	0.758	
11	0.276	0.360	-0.584	1.150	0.731	0.000	0.552	0.944	-0.053	*
12	-0.623	0.000	-0.785	-0.484	0.864	0.000	0.752	0.992	0.750	
13	-0.031	0.805	-0.245	0.214	0.587	0.000	0.500	0.684	0.413	*
14	-0.058	0.643	-0.380	0.279	0.647	0.000	0.538	0.741	0.370	*
15	-0.475	0.000	-0.660	-0.281	0.888	0.000	0.785	0.983	0.581	

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 4, 11, 13 and 14. Although the estimations for 200 m showed better, on average, Adjusted R^2's than the 150 m layer, some clusters that had significant results in the 150 m layer, like 4 for example, were in this case non-significant.

The following results were obtained for the differences model, also for the 200 m layer:

Table 43: Summary of Results, Disaggregated Analysis, Instant Games, 200 m, Differences Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.526	0.277	0.274	0.4045	86.277	0
2	0.573	0.328	0.326	0.3474	159.714	0
4	0.555	0.308	0.3	0.3754	38.939	0
35	0.596	0.355	0.35	0.3995	70.13	0
6	0.662	0.438	0.436	0.3181	174.056	0
7	0.578	0.334	0.328	0.3788	51.729	0
8	0.66	0.435	0.431	0.3009	102.854	0
9	0.674	0.455	0.433	0.3062	20.854	0
10	0.545	0.297	0.293	0.4633	70.67	0
11	0.73	0.533	0.52	0.2606	41.599	0
12	0.556	0.309	0.307	0.3830	135.034	0
13	0.572	0.328	0.324	0.3849	87.181	0
14	0.497	0.247	0.243	0.4216	65.533	0
15	0.762	0.581	0.578	0.2758	167.929	0

As seen in the table, the differences model delivered Adjusted R^2's in the range [0.18, 0.58] for all the clusters, and in every case, the F-tests indicated that the regressions were significant. Adjusted R^2's were slightly higher than in the case of the linear models, and higher also than in the estimations for 150 m. The average Adjusted R^2 was 0.37. The following table shows the results obtained for the parameters:

Table 44: Summary of Results with Coefficients, Disaggregated Analysis, Instant Games, 200 m, Differences Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.767	0.000	-1.128	-0.511	0.767	0.000	0.652	0.873	0.984	
2	-0.527	0.000	-0.816	-0.196	0.835	0.000	0.750	0.928	0.686	
4	-0.199	0.231	-0.688	0.337	0.669	0.000	0.530	0.796	0.487	*
35	-0.607	0.001	-0.916	-0.357	0.991	0.000	0.820	1.157	0.611	
6	-0.427	0.013	-0.849	-0.137	0.981	0.000	0.892	1.075	0.443	
7	-0.758	0.000	-1.036	-0.492	0.937	0.000	0.753	1.099	0.813	
8	-0.662	0.001	-1.225	-0.185	0.875	0.000	0.770	0.990	0.771	
9	-0.785	0.001	-1.159	-0.382	0.944	0.000	0.685	1.185	0.866	
10	-0.727	0.003	-0.958	-0.530	0.937	0.000	0.742	1.179	0.750	
11	-0.077	0.830	-0.763	0.509	0.776	0.000	0.582	0.958	0.266	*
12	-0.681	0.000	-0.827	-0.533	0.858	0.000	0.733	0.998	0.820	
13	-0.057	0.637	-0.316	0.215	0.627	0.000	0.540	0.714	0.402	*
14	-0.163	0.209	-0.501	0.194	0.659	0.000	0.549	0.785	0.484	*
15	-0.522	0.000	-0.700	-0.333	0.935	0.000	0.838	1.030	0.575	

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 4, 11, 13 and 14, the same ones as in the linear case.

300 m Layer

The following results were obtained for the estimation of the linear models, for the 300 m layer:

Table 45: Summary of Results, Disaggregated Analysis, Instant Games, 300 m, Linear Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.431	0.186	0.181	0.4618	33.288	0
2	0.603	0.363	0.359	0.3024	92.906	0
4	0.542	0.293	0.284	0.3423	32.182	0
35	0.623	0.388	0.381	0.3738	57.724	0
6	0.6	0.36	0.355	0.3221	75.149	0
7	0.676	0.457	0.45	0.2740	63.892	0
8	0.714	0.51	0.503	0.2927	78	0
9	0.901	0.811	0.795	0.2058	49.345	0
10	0.617	0.38	0.373	0.3738	51.545	0
11	0.848	0.719	0.697	0.2242	32.063	0
12	0.569	0.324	0.321	0.3866	97.642	0
13	0.555	0.308	0.301	0.4157	44.03	0
14	0.477	0.228	0.222	0.4165	38.375	0
15	0.79	0.624	0.621	0.2443	162.079	0

As seen in the table, the linear model delivered Adjusted R^2's in the range [0.18, 0.8], which is the highest of any layer for instant games. Removing the observation with the lowest value (cluster 1), the next lowest Adjusted R^2 is 0.23, and the average is 0.44. All the F-tests indicated that the regressions delivered significant results. The following table shows the results obtained for the parameters:

Table 46: Summary of Results with Coefficients, Disaggregated Analysis, Instant Games, 300 m, Linear Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.130	0.457	-0.480	0.214	0.703	0.000	0.559	0.837	0.406	*
2	-0.690	0.000	-0.876	-0.510	0.781	0.000	0.655	0.909	0.897	
4	-0.021	0.848	-0.361	0.257	0.615	0.000	0.459	0.767	0.385	*
35	-0.919	0.000	-1.390	-0.488	1.035	0.000	0.857	1.222	0.872	
6	-0.325	0.011	-0.635	-0.022	0.874	0.000	0.741	1.004	0.448	*
7	-0.822	0.000	-1.102	-0.533	0.939	0.000	0.775	1.092	0.859	
8	-0.883	0.000	-1.116	-0.676	0.847	0.000	0.731	0.963	1.004	
9	-0.938	0.000	-1.360	-0.582	1.020	0.000	0.802	1.302	0.918	
10	-0.528	0.029	-0.713	-0.254	0.923	0.000	0.772	1.104	0.600	
11	-1.355	0.001	-2.127	-0.856	0.849	0.000	0.594	1.086	1.405	
12	-0.617	0.000	-0.817	-0.412	0.974	0.000	0.813	1.154	0.639	
13	-0.419	0.006	-0.858	-0.048	0.676	0.000	0.557	0.803	0.686	*
14	-0.205	0.073	-0.463	0.110	0.666	0.000	0.503	0.815	0.510	*
15	-0.454	0.000	-0.696	-0.184	0.938	0.000	0.848	1.014	0.523	

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 1, 4, 6, 13 and 14.

The following results were obtained for the differences model, also for the 300 m layer:

Table 47: Summary of Results, Disaggregated Analysis, Instant Games, 300 m, Differences Model

Id	R	R^2	Adjusted R^2	Std. Error of the Estimate	F	Sig.
1	0.478	0.229	0.223	0.4356	38.682	0
2	0.613	0.376	0.372	0.3005	89.44	0
4	0.573	0.329	0.319	0.3244	34.531	0
35	0.649	0.421	0.414	0.3586	60.078	0
6	0.648	0.419	0.415	0.2912	87.8	0
7	0.697	0.486	0.478	0.2653	65.199	0
8	0.732	0.536	0.529	0.2856	78.46	0
9	0.899	0.809	0.791	0.2129	44.457	0
10	0.678	0.46	0.453	0.3376	65.591	0
11	0.848	0.718	0.693	0.1877	28.065	0
12	0.573	0.328	0.325	0.3848	91.474	0
13	0.641	0.411	0.405	0.3530	62.163	0
14	0.464	0.215	0.208	0.4203	32.051	0
15	0.802	0.644	0.64	0.2417	160.849	0

As seen in the table, the differences model delivered Adjusted R^{2}'s in the range [0.21, 0.79], which is very similar to what was seen in the case of the linear model. The average Adjusted R^2 is 0.45. All the F-tests indicated that the regressions delivered significant results. The following table shows the results obtained for the parameters:

Table 48: Summary of Results with Coefficients, Disaggregated Analysis, Instant Games, 300 m, Differences Model

	Active Normalized (cannibalization coefficient)				Seasonal Factor				Constant	Mark
Id	B	Sig.	95% BT LL	95% BT UP	B	Sig.	95% BT LL	95% BT LL	B	
1	-0.134	0.435	-0.467	0.211	0.725	0.000	0.592	0.868	0.388	*
2	-0.695	0.000	-0.887	-0.504	0.771	0.000	0.649	0.900	0.916	
4	0.059	0.575	-0.327	0.359	0.637	0.000	0.478	0.779	0.285	*
35	-0.794	0.000	-1.154	-0.483	1.044	0.000	0.840	1.227	0.750	
6	-0.292	0.015	-0.570	0.031	0.884	0.000	0.746	0.995	0.402	*
7	-0.783	0.000	-1.096	-0.502	0.952	0.000	0.796	1.111	0.815	
8	-0.939	0.000	-1.169	-0.738	0.852	0.000	0.735	0.971	1.061	
9	-0.956	0.000	-1.461	-0.665	1.022	0.000	0.772	1.375	0.936	
10	-0.538	0.015	-0.714	-0.317	0.960	0.000	0.808	1.148	0.558	
11	-1.145	0.001	-1.635	-0.693	0.739	0.000	0.539	0.906	1.347	
12	-0.635	0.000	-0.819	-0.417	0.959	0.000	0.806	1.162	0.678	
13	-0.314	0.020	-0.568	-0.020	0.692	0.000	0.552	0.805	0.568	*
14	-0.219	0.067	-0.473	0.079	0.660	0.000	0.509	0.804	0.541	*
15	-0.477	0.000	-0.754	-0.222	0.953	0.000	0.866	1.035	0.524	

The clusters with a "*" in the column "Mark" were those in which the estimation of the cannibalization parameter was either non-significant at the 95% level and/or the bootstrap showed a parameter that was highly unstable. This was the case of clusters 1, 4, 6, 13 and 14, the same as in the linear model.

Summary

The same trade-offs between geographic precision and statistical reliability observed in the case of lottery games are also present here, but they become more critical as the average Adjusted R^2's are significantly lower now. The trade-off between demand clusters and aggregated estimation (more observations in the case of aggregated estimation, but disregarding demand differences) is also present.

The following table summarizes the expected errors (same definition as the one used for lottery games) obtained for the cannibalization coefficients, estimated using the linear functions, for the demand clusters and the aggregated estimation:

Table 49: Summary of Errors, Linear Model, Instant Games

	ID	150 m	200 m	300 m
Demand Clusters	1	36%	39%	267%
	2	34%	59%	27%
	4	64%	228%	1471%
	35	38%	46%	49%
	6	73%	69%	94%
	7	1945%	35%	35%
	8	39%	65%	25%
	9	59%	48%	41%
	10	25%	28%	43%
	11	303%	314%	47%
	12	31%	24%	33%
	13	545%	740%	97%
	14	67%	568%	140%
	15	88%	40%	56%
	Average[13]	**48%**	**45%**	**40%**
Aggregated		18%	22%	27%

The average error decreases as the radius increases, but, contrary to what was observed in the previous product category, the aggregated error increases as the radius increases. In any case, the expected errors of the aggregated estimation are on average half of the ones obtained with the estimation at the demand cluster level.

[13] The average excludes in every layer the clusters where the coefficients were not significant. In the 150 m layer, clusters 6, 7, 11 and 13 were excluded; in the 200 m layer, clusters 4, 11, 13 and 14 were excluded and in the 300 m layer, clusters 1, 4, 6, 13 and 14 were excluded.

The 150 m layer was selected for the location assessment presented in the following chapter for instant games: it offers the highest geographical precision, has the same number of demand clusters with expected errors below 31% as the other two layers and has the lowest aggregated error. Once the 150 m layer was selected, it was necessary to define a procedure to choose the set of coefficients for each demand cluster, considering that in some of them the expected precision by using the parameters estimated at the level of demand clusters was too low. The procedure was as follows:

- If the expected error at the level of the demand cluster based on the cannibalization coefficients was equal, less or marginally superior to 31%, then the set of coefficients estimated for the demand cluster was used.
- If the expected error for a demand cluster based on the cannibalization coefficients was significantly superior to 31%, then the expected error of the "closest" demand cluster, if there was a cluster close enough in terms of demand characteristics, was examined. If it was less than 31%, then the coefficients for the closest demand cluster were used.
- In any other case, the coefficients of the aggregated estimation were used.

The following table presents a summary of this process:

Table 50: Selection of the Coefficients per Cluster, Instant Games

ID	Expected Error	Less than 31%?	Closest cluster	Error closest cluster	Closest Less than 31%?	Cluster of the final coefficients to use
1	36%	No	2	34%	No	Aggregated
2	34%	No	1	36%	No	Aggregated
4	64%	No	35	38%	No	Aggregated
35	38%	No	4	64%	No	Aggregated
6	73%	No	7	1945%	No	Aggregated
7	1945%	No	6	73%	No	Aggregated
8	39%	No	10	25%	Yes	10
9	59%	No	none close enough			Aggregated
10	25%	Yes				10
11	303%	No	10	25%	Yes	10
12	31%	Yes				12
13	545%	No	14	67%	No	Aggregated
14	67%	No	15	88%	No	Aggregated
15	88%	No	14	67%	No	Aggregated

Linear models were also preferred in this case because the linear functional forms obtained have properties that will make the location assessment presented in the next chapter easier to calculate. The differences model delivered results that in most cases were comparable to the ones obtained using the linear model, and in cases when there was a difference, the increment in precision and Adjusted R^2 was marginal.

Chapter 5: Growth Assessment and Recommendations

Based on the cannibalization functions estimated in the previous chapter for every product category, the purpose of this chapter is to provide recommendations regarding a potential growth strategy. Accordingly, as it was shown, the distribution network is starting to evidence signs of saturation, as increases in the number of stores have effectively caused existing stores to decrease their average sales level. The implication of this finding is that the proportion of new sales obtained as the result of store openings is actually decreasing, reducing the efficiency (in sales per store or square feet, for example) of the network. The findings also suggest, as will be shown in this chapter, that beyond a certain threshold for each geocluster, every new store added to the network just redistributes existing sales, in zero-sum games, without bringing any new sales to the company.

The coefficients obtained for cannibalization also show that the situation is not the same for both product categories: lottery games, demanded by consumers in a more planned fashion, tend to be more subject to cannibalization than instant games, whose demand is mostly driven by impulse. The implication of this finding is, therefore, that instant games can support, and in fact require, a more extensive distribution network.

Based on these analyses it was found that certain areas of Santiago are already saturated for one or both product categories, but that there are others where the distribution network can still support additional stores that would bring new sales. This chapter shows the procedure followed to assess the areas' potential, as well as the main results obtained.

The Sales Function

In Chapter 4, the following relationship was defined, for each product category and geoclustering layer:[14]

$$t_{ij} = \alpha \cdot a_{ij} + \beta \cdot f_i + C$$

where:

[14] This is the linear form of the relationship. As explained at the end of Chapter 4, the linear form was preferred over the differences model because it delivered similar results, but it was easier to manipulate for the purposes of this chapter.

T_{ij} : Average size of the stores of the geo cluster *j* at the time *i*

A_{ij} : Number of active stores in the geo cluster *j* at the time *i*

$\overline{T_j}$: Average size of the stores of the geo cluster *j*

$\overline{A_j}$: Average number of active stores at the geo cluster j

$t_{ij} = \dfrac{T_{ij}}{\overline{T_j}}$: Normalized size of the stores of the geo cluster j at the time i

$a_{ij} = \dfrac{A_{ij}}{\overline{A_j}}$: Normalized number of active stores at the geo cluster j at the time i

f_i : Seasonal factor for time *i*

The parameters for this function were estimated for each product category, demand cluster and geoclustering layer. The final set of coefficients, i.e., which geocluster layer to use in every case, was chosen based on the width obtained for the range of α, the cannibalization coefficient, through a 500-sample bootstrap.

By replacing the original variables in the linear function and then multiplying it by the average size of the clients of each geocluster, it is possible to obtain the average size of the clients of geocluster *j* during month *i*:

$$\frac{T_{ij}}{\overline{T}_j} = \alpha \cdot \frac{A_{ij}}{\overline{A}_j} + \beta \cdot f_i + C \Big/ \times \overline{T}_j$$

Multiplying then by the number of active stores in the month *i* (A_{ij}), it is possible to obtain the month's total sales for geocluster *j*:

$$T_{ij} = \alpha \cdot \frac{\overline{T}_j}{\overline{A}_j} \cdot A_{ij} + \beta \cdot f_i \cdot \overline{T}_j + C \cdot \overline{T}_j \Big/ \times A_{ij}$$

$$T_{ij} \cdot A_{ij} = \alpha \cdot \frac{\overline{T}_j}{\overline{A}_j} \cdot A^2{}_{ij} + \beta \cdot f_i \cdot \overline{T}_j \cdot A_{ij} + C \cdot \overline{T}_j \cdot A_{ij}$$

By deriving the total sales in the geocluster with respect to the number of active stores, and then making it equal to zero, it is possible to obtain the number of active stores that maximizes the sales in the geocluster:

$$Sales_{ij} = T_{ij} \cdot A_{ij} = \alpha \cdot \frac{\overline{T}_j}{\overline{A}_j} \cdot A^2{}_{ij} + \beta \cdot f_i \cdot \overline{T}_j \cdot A_{ij} + C \cdot \overline{T}_j \cdot A_{ij} \Big/ \times \frac{\partial}{\partial A_{ij}}$$

$$\frac{\partial Sales_{ij}}{\partial A_{ij}} = 2 \cdot \alpha \cdot \frac{\overline{T}_j}{\overline{A}_j} \cdot A_{ij} + \beta \cdot f_i \cdot \overline{T}_j + C \cdot \overline{T}_j = 0$$

$$A_{ij}^{Saturation} = \frac{-\beta \cdot f_i \cdot \overline{T}_j - C \cdot \overline{T}_j}{2 \cdot \alpha \cdot \frac{\overline{T}_j}{\overline{A}_j}}$$

$A_{ij}^{Saturation}$ is the number of active stores that "saturates" geocluster *i* during month *j*. Stores beyond this saturation threshold represent a theoretical zero-sum game.

The function obtained for the saturation number of stores shows behaviors aligned with logic:

- As the cannibalization coefficient, α, decreases, the number of stores at saturation increases. As the coefficient tends to zero (no cannibalization), the number at saturation tends to infinity. This makes sense as a product category without cannibalization will never saturate its distribution network, and every new store will provide additional sales (hypothetical case of infinite demand, or when demand >> current sales level).
- The saturation number of stores depends on the season: in higher seasons, represented by higher values for f_i, the seasonal factor, geoclusters will be able to support more stores, but when the season is low, the saturation point will also be lower. This represents a challenge for the design of the distribution network, as the optimal number of stores will vary month to month.

The following sections show calculations for the saturation number of stores for each geocluster and product category, assuming a seasonality factor of 1 (an average month of the year). A saturation gap for each geocluster and product category was calculated as the difference between the saturation number of stores and the number of active stores for the last month with available information.

Saturation and New Store Openings for Lottery Games

Using the function for the saturation number of stores, the following figure shows a distribution of the saturation gaps for lottery games, evaluated using the coefficients selected for each cluster, based on the criteria shown in Chapter 4:

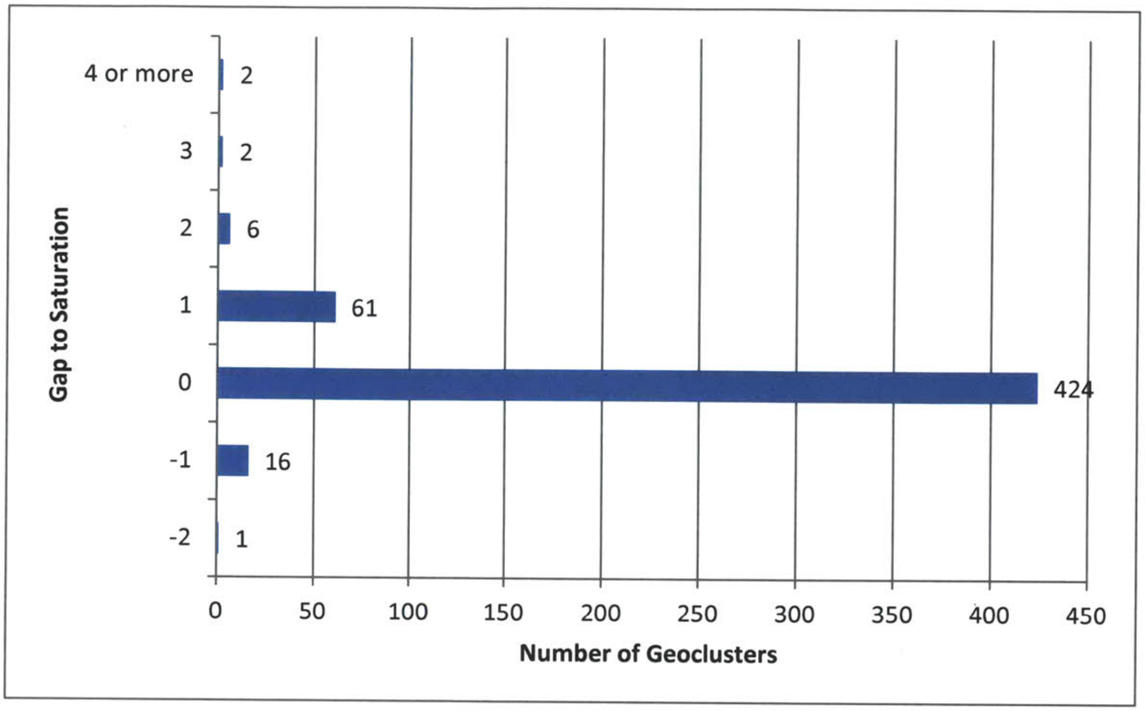

Figure 32: Distribution of the Gap to Saturation, Lottery Games

As seen, most of the geoclusters are currently at or beyond their saturation level (83% of them). These geoclusters should not be the target for new store openings. There are 71 geoclusters, however, that still could support more stores; of these geoclusters, only 10 would support more than one store.

The previous graph was calculated using the cannibalization parameters estimated through least squares; using the 500-sample bootstrap, however, lower and upper 95% limits were calculated for the parameters. In this regard, the following figure shows a sensitivity analysis of the gap to saturation using the upper and lower limits for the cannibalization parameters:

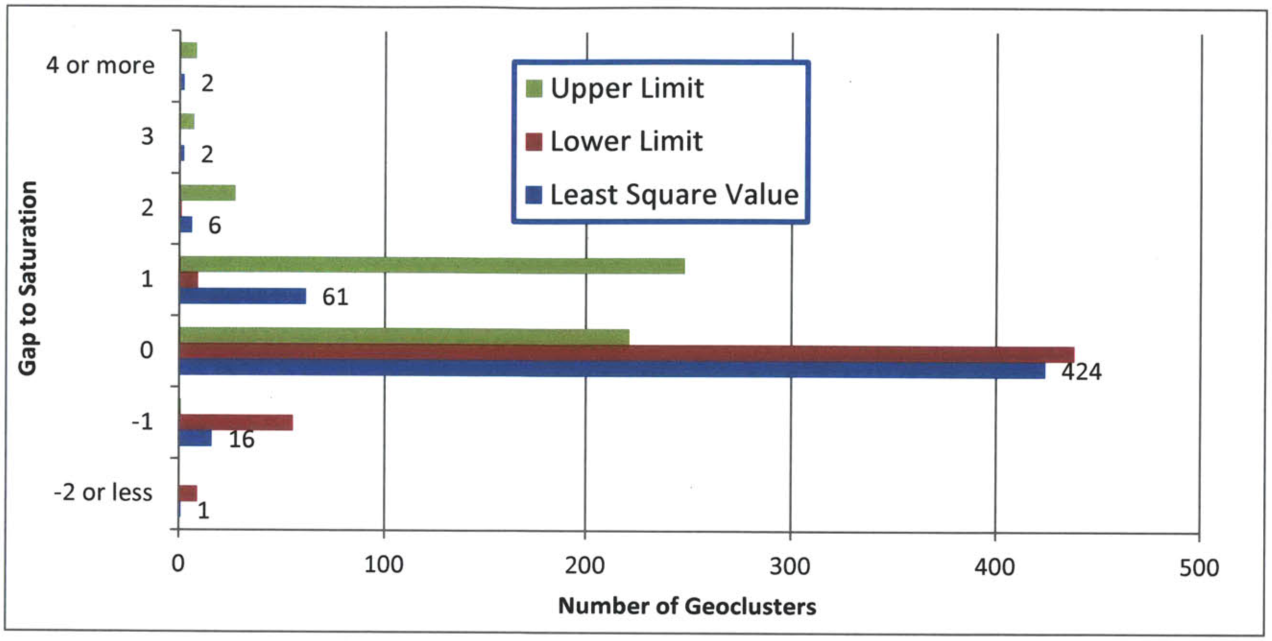

Figure 33: Distribution of the Gap to Saturation by Parameter Selected, Lottery Games

As the values used for the cannibalization coefficient vary, so does the number of stores to saturation. Accordingly, with more a "conservative" estimation, based on the lower 95% bootstrap limit for α (a negative number), as little as 10 geoclusters, out of 512, still have room for one additional store. If, on the other hand, the upper limit is selected, the number of geoclusters with room for additional stores grows up to 290, with 42 of them with room for more than one additional store.

As seen in the graph, geoclusters have different gaps according to the coefficients used. If the company wants to increase sales through new store openings, according to this model, it should select geoclusters where the gap to saturation is greater, as they are the ones with the bigger opportunities. The problem, however, is that the gap varies with the parameters selected. In this regard, the following table summarizes the combination of gaps according to the parameters. The first column, "Combination" is a label created to refer later to the combination; the second column, "Gap low α," is the gap to saturation obtained with the bootstrap lower limit; "Gap α" is the gap with the least square parameter; "Gap high α" is the gap with the bootstrap upper limit, and "N° of Cases" is the number of geoclusters in that combination. In this way, combination 1 in the first row for example, indicates that there are 2 geoclusters where the gap using the bootstrap lower limit for α is -3 or less, the gap using least squares is -1 or less and the gap using the bootstrap lower limit is -1 or less. Geoclusters in combination 1, therefore, are highly saturated:

Table 51: Summary of Cases to Saturation, Lottery Games

Combination	Gap low α	Gap α	Gap high α	N° of Cases
1	-2 or less	-2 or less	0 or less	3
2	-2	-1	0	2
3	-2	-1	1	2
4	-2	0	3	2
5	-1	-1	0	10
6	-1	0	0	33
7	-1	0	1	6
8	-1	0	2	3
9	-1	1	2	1
10	-1	1	3	1
11	-1	12	28	1
12	0	0	0	175
13	0	0	1	204
14	0	0	2	1
15	0	1	1	34
16	0	1	2	18
17	0	1	3	2
18	0	2	4	3
19	0	3	6	1
20	1	1	1	2
21	1	1	2	3
22	1 or more	2 or more	3 or more	5

Which combinations to select to grow depend on the level of aggressiveness of the pursued growth strategy. For example, a highly conservative strategy would select as geoclusters to increase the number of stores combinations 20, 21 and 22, where the gap, regardless of the coefficients, is always positive. In contrast, even with a highly aggressive strategy, combination 1 should never be selected, as almost for sure new stores will only bring cannibalization to the network. Based on this table, the following classification is proposed for the geoclusters in terms of the risk associated with the growth strategy:

- **Go, priority 1**. Combinations 20 to 22: low risk of severe cannibalization when incrementing the number of stores. An additional store most likely will provide additional sales. Recommended to explore.
- **Analyze, priority 2**. Combinations 15 to 19 and 11: moderate risk of severe cannibalization when incrementing the number of stores. An additional store has good chances of producing new sales. Recommended to explore.
- **Analyze thoroughly, priority 3**. Combinations 12 to 14: moderate to high risk of severe cannibalization when incrementing the number of stores. An additional store has moderate chances of just redistributing sales. Not recommended to explore.

- **Do not go unless necessary, priority 4**. Combinations 7 to 10: high risk of severe cannibalization when incrementing the number of stores. An additional store has high chances of just redistributing sales. Not recommended to explore.
- **Do not go, priority 5**. Combinations 2 to 6: very high risk of severe cannibalization when incrementing the number of stores. An additional store has very high chances of just redistributing sales. Not recommended to explore.
- **Do not go**. Combination 1: almost for certain new stores will bring only cannibalization, redistributing sales. Do not increment the number of stores.

The following map shows the geoclusters shaded according to the previous classification:

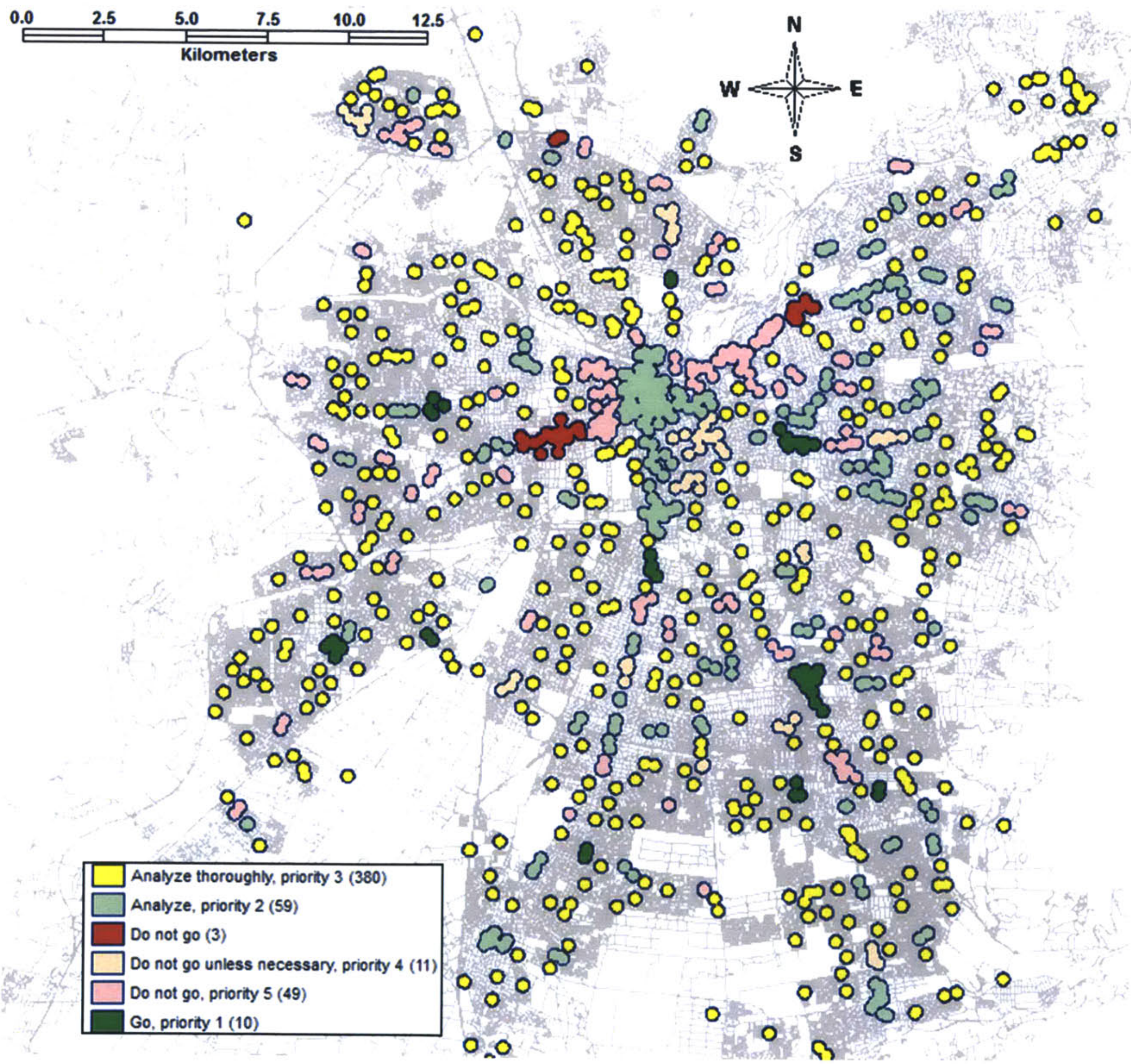

Figure 34: Geoclusters Prioritized for New Store Openings, Lottery Games

Accordingly, when looking for new stores, the company should prospect first geoclusters in the following order:

- First green ("Go, priority 1").
- Then light green ("Analyze, priority 2").

If additional stores were still required, the company should carefully analyze yellow geoclusters ("Analyze thoroughly, priority 3"), orange geoclusters (Do not go unless necessary, priority 4), and then pink geoclusters ("Do not go, priority 4"). Red geoclusters, as it was mentioned, should not even be considered.

Within each color code, geoclusters could be prioritized, according to the "Combination id" previously introduced (in inverse order, starting in green geoclusters, for example, by combination 22, then 21, and so on) or by the gap to saturation (geoclusters with bigger gaps should go first). The issue associated with using the gap to saturation is that they vary according to the set of coefficients selected to calculate them (as shown in the previous table), which could eventually deliver different prioritizations for geoclusters depending on the coefficients chosen.

Saturation and New Store Openings for Instant Games

Similar to the previous case, using the function for the saturation number of stores, the following figure shows the distribution of the saturation gaps for instant games, evaluated using the coefficients selected for each cluster, based on the criteria shown in Chapter 4:

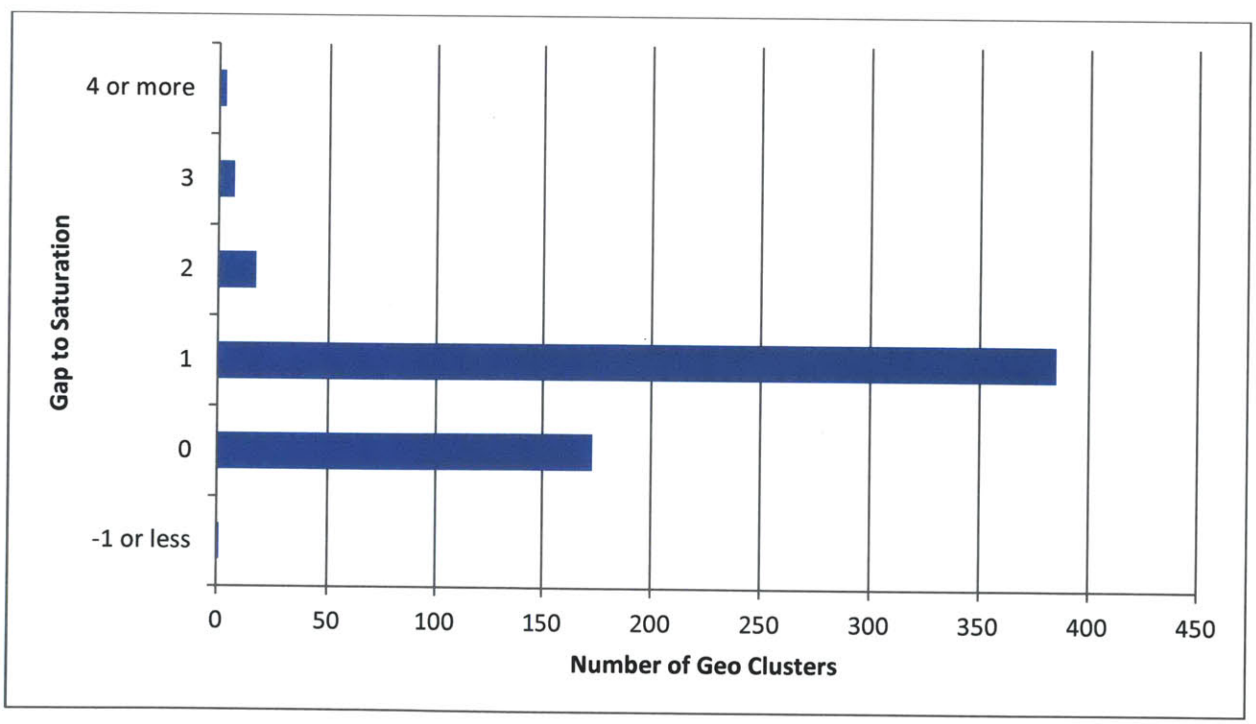

Figure 35: Distribution of the Gap to Saturation, Instant Games

Unlike what happened with lottery games, in instant games most of the geoclusters have not reached their saturation points yet. This was a foreseeable situation, as the number of stores selling instant games is almost the same as those selling lottery games, but the cannibalization coefficients are significantly lower. Accordingly, more than 70% of the geoclusters can still support (mostly) one or more additional stores

The previous graph was calculated using the cannibalization parameters estimated through least squares; using the 500-sample bootstrap, however, lower and upper 95% limits were calculated for the parameters. In this regard, the following figure shows a sensitivity analysis of the gap to saturation using the upper and lower limits for the cannibalization parameters:

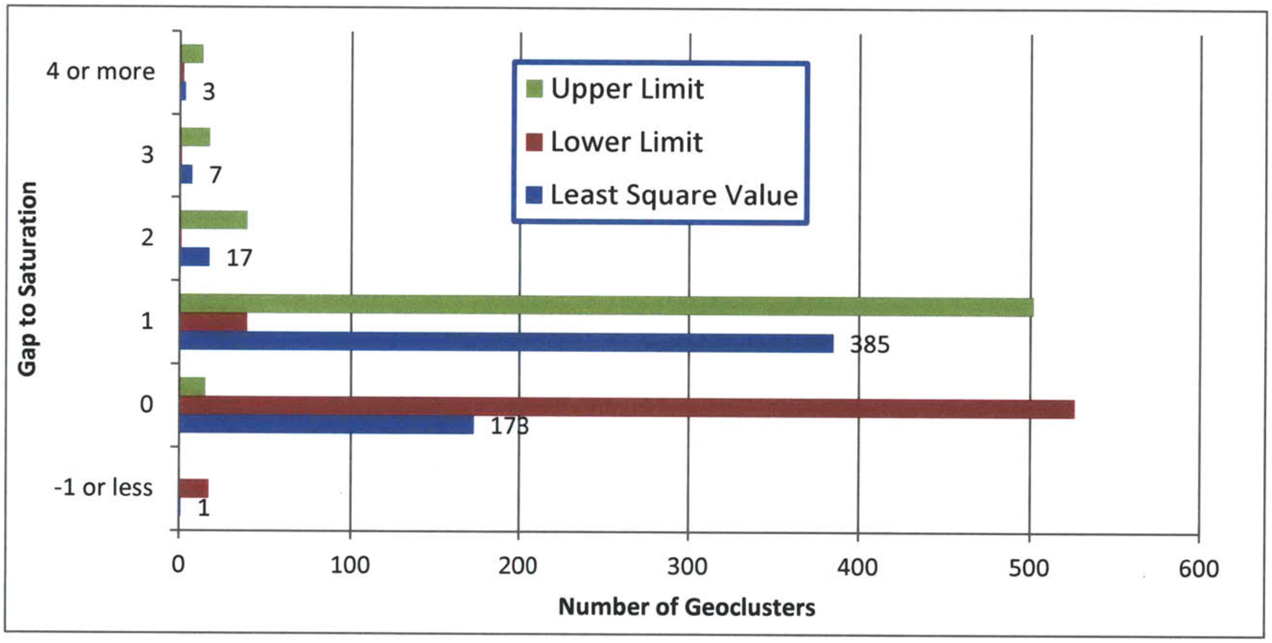

Figure 36: Distribution of the Gap to Saturation by Parameter Selected, Instant Games

As the values used for the cannibalization coefficient vary, so does the number of stores to saturation. Accordingly, with more a "conservative" estimation, based on the lower 95% bootstrap limit for α (a negative number), as little as 43 geoclusters, out of 586, still have room for one additional store. If, on the other hand, the upper limit is selected, the number of geoclusters with room for additional stores grows up to 571, with 69 of them with room for more than one additional store.

As seen in the distribution graph, geoclusters have different gaps according to the coefficients used. If the company wants to increase sales through new store openings, according to this model, it should select geoclusters where the gap to saturation is greater, as they are the ones with the greatest opportunities. The problem, as in the previous case, is that gaps vary with the parameters selected. In this regard, the following table summarizes the combination of gaps according to the set of parameters. The first column, "Combination" is a label created to refer to the combination later, and also serves as a ranking (in inverse order) of the level of attractiveness of the geoclusters; the second column, "Gap low α" is the gap to saturation obtained with the bootstrap lower limit; "Gap α" is the gap with the least square parameter; "Gap high α" is the gap with the bootstrap upper limit and "N° of Cases" is the number of geoclusters in that combination. In this way, combination 1 in the first row for example, indicates that there are 15 geoclusters where the gap using the bootstrap lower limit for α is 0 or less, the gap using least squares is 0 or less and the gap using the bootstrap lower limit is 0 or less. Geoclusters in combination 1, therefore, are highly saturated:

Table 52: Summary of Cases to Saturation, Instant Games

Combination	Gap low α	Gap α	Gap high α	N° of Cases
1	0 or less	0 or less	0 or less	15
2	-3	-1	2	1
3	-2	0	2	1
4	-2	2	9	1
5	-1	0	1	5
6	-1	0	2	2
7	0	0	1	149
8	0	0	2	1
9	0	1	1	346
10	0	1	2	16
11	0	1	3	4
12	0	2	4	1
13	0	3	7	1
14	1	1	1	2
15	1	1	2	17
16	1	2	2	1
17	1	2	3	13
18	1	2	4	1
19	1 or more	3 or more	4 or more	9

Which combinations to select for growing depend on the level of aggressiveness of the pursued growth strategy. For example, a highly conservative strategy would select as geoclusters to increase the number of stores only combinations 14 through 19 where the gap, regardless of the coefficients, is always positive. In contrast, a highly aggressive strategy should never select combination 1, where almost for sure new stores will only bring cannibalization to the network. Based on this table, the following classification is proposed for the geoclusters in terms of the risk associated with the growth strategy:

- **Go, priority 1**. Combinations 14 to 19: low risk of severe cannibalization when incrementing the number of stores. An additional store most likely will provide new sales. Recommended to explore.
- **Analyze, priority 2**. Combinations 9 to 13: moderate risk of severe cannibalization when incrementing the number of stores. An additional store has good chances of producing new sales. Recommended to explore.

- **Analyze thoroughly, priority 3**. Combinations 7 and 8: high risk of severe cannibalization when incrementing the number of stores. An additional store has high chances of just redistributing sales. Not recommended to explore.
- **Do not go, priority 4**. Combination 2 to 6: very high risk of severe cannibalization when incrementing the number of stores. An additional store has very high chances of just redistributing sales. Not recommended to explore.
- **Do not go**. Combination 1: almost for certain new stores will bring only cannibalization, redistributing sales. Do not increment the number of stores.

The following map shows the geoclusters shaded according to the previous classification:

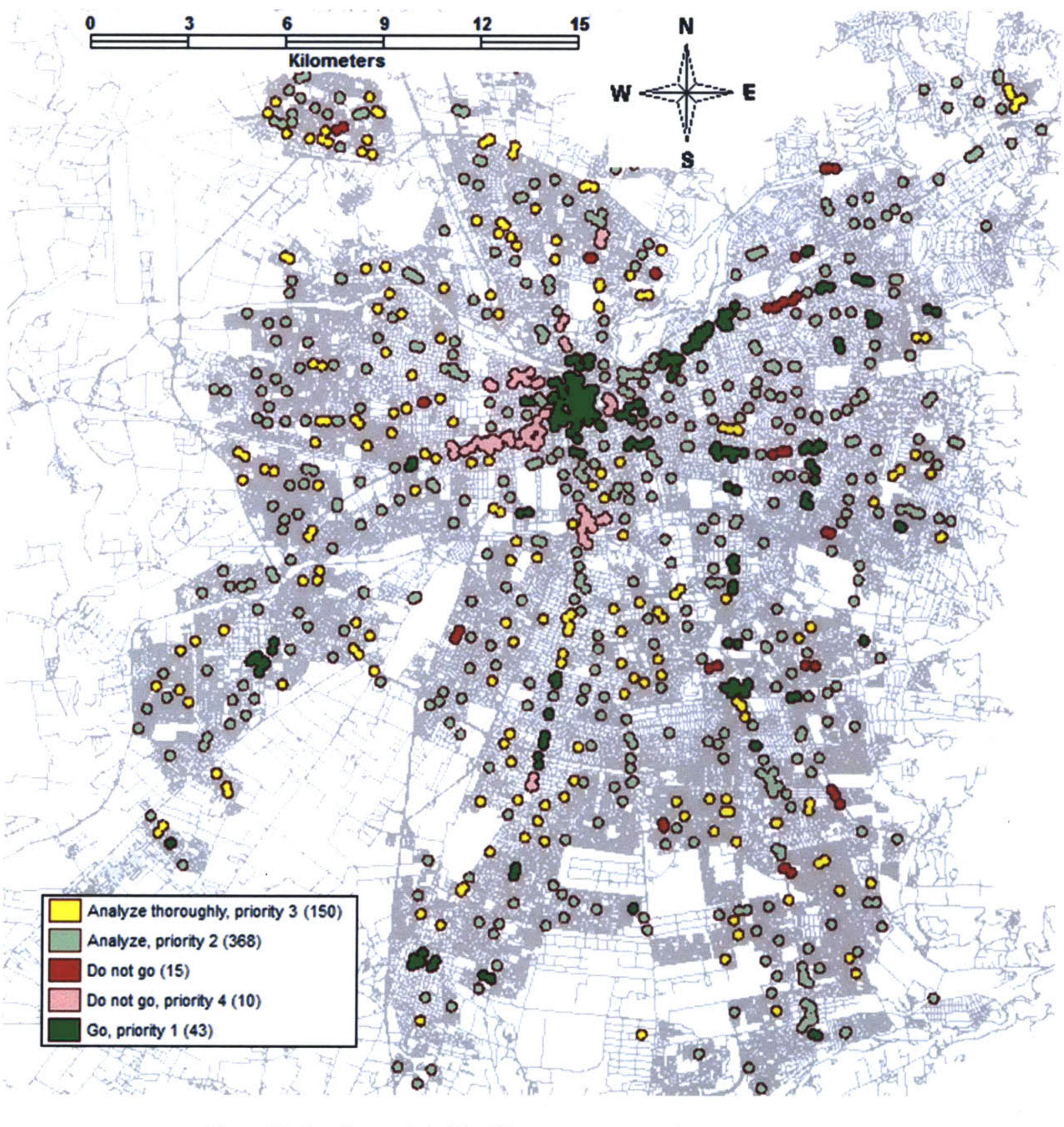

Figure 37: Geoclusters Prioritized for New Store Openings, Instant Games

Accordingly, when looking for new stores, the Company should prospect first geoclusters in the following order:

- First green ("Go, priority 1").
- Then light green ("Analyze, priority 2").

If additional stores were still required, the company should carefully analyze yellow geoclusters ("Analyze thoroughly, priority 3") and then pink geoclusters ("Do not go, priority 4"). Red geoclusters, as mentioned, should not even be considered.

Within each color code, geoclusters could be prioritized according to the "Combination id" previously introduced (in inverse order, starting in green geoclusters for example, by combination 19, then 18, and so on) or by the gap to saturation (geoclusters with bigger gaps should go first). The issue associated with using the gap to saturation is that they vary according to the set of coefficients selected to calculate them (as shown in the previous table), which could eventually deliver different prioritizations for geoclusters depending on the coefficients chosen.

Additional Considerations and Further Areas of Research

As mentioned at the beginning of this chapter, all the calculations for the number of stores to saturation were made considering a seasonal factor of 1, an average month during the year. Seasonality, however, plays an important role in the model, as the parameter for seasonal factors always showed significant results in the regressions. Accordingly, the number of stores to saturation is expected to vary with the season: during high seasons, when the jackpot is accumulated for example, geoclusters will momentarily support more stores until reaching saturation, while during lower seasons, saturation thresholds will be lower. To illustrate this effect, the following table shows a comparison of the total number of stores to saturation for each product category, considering the extreme values observed for the seasonal factors:

Table 53: Gaps to Saturation with Extreme Seasonal Factors

Product	Gap Average Month	Gap in the lowest month	Gap in the Highest Month
Instant Games	473	-57	807
Lottery Games	77	-114	1,257
Total	550	-171	2,064

As expected, the need for stores changes according to the season. For the month with the lowest demand, the network of stores is already saturated for both product categories, while for the month with the highest demand, the network could accommodate more than 800 additional stores for instant games, and almost 1,300 for lottery games. The ranges (the difference between the saturation numbers of stores for the highest and lowest month) also vary among product categories, as lottery games show a sales behavior much more influenced by seasonality (due to jackpot accumulations) than instant games.

This level of variability calls for a flexible design of the store network. De Neufville et al. (2011) describe a methodology to use "real options" in engineering projects where variability and uncertainty are important factors. In this case, flexibility could be implemented by keeping the current number of stores fixed, for example, and managing demand peaks using some sort of "mobile" infrastructure like kiosks or modular terminals that could be located during peak season in selected out-of-network stores, or through incentives to consumers buy online. Although beyond the scope of this thesis, policies to mitigate the effects of seasonality in the network, like the ones previously discussed, and mechanisms to evaluate them, would be useful in this context.

Another aspect that could be included in the analysis to calculate the optimal number of stores per geocluster are the fixed costs associated with setting additional stores (fundamentally, installation costs and an increase in the cost to serve the stores) and the average profit margin of sales. Accordingly, if fixed costs are significant, the optimal number of stores per geocluster would be lower than the saturation number, as new stores not only have to bring additional sales, but also have to cover the increment in fixed costs. If, on the other hand, fixed costs are non-significant, having exactly the saturation number of stores in each geocluster maximizes the company's profits. In this case, fixed costs are relevant, as the company invests to set new stores and incurs fixed expenses to serve them; the saturation number of stores should be then taken as a reference not to go beyond, but not as an optimum.

Another important aspect to consider is the presence of "commonality" in the sense that in some cases, the same geoclusters present opportunities to increment the number of stores for both product categories. The following table shows a summary of the geoclusters, classified according to the saturation categorization proposed for each product at the end of the previous sections:

Table 54: Instant Game Prioritization vs. Lottery Games Prioritization

		Lottery Games					
		Go, priority 1	Analyze, priority 2	Analyze thoroughly, priority 3	Do not go unless necessary, priority 4	Do not go, priority 5	Do not go
Instant Games	Go, priority 1	4	25	5	5	7	1
	Analyze, priority 2	4	44	259	17	42	2
	Analyze thoroughly, priority 3	8	21	100	4	14	3
	Do not go, priority 4		3	2	1	5	2
	Do not go		2	5		8	

As seen, there are geoclusters that would be targets for both product categories and others that, in contrast, are already saturated also in both. Accordingly, the following map shows geoclusters categorized as follows:

- High Priority Common (green): geoclusters that were priority 1 or 2 for both product categories.
- Low Priority Common (red): geoclusters with the lowest priority and the "Do not go" label for both product categories.
- All the rest (yellow): all the rest.

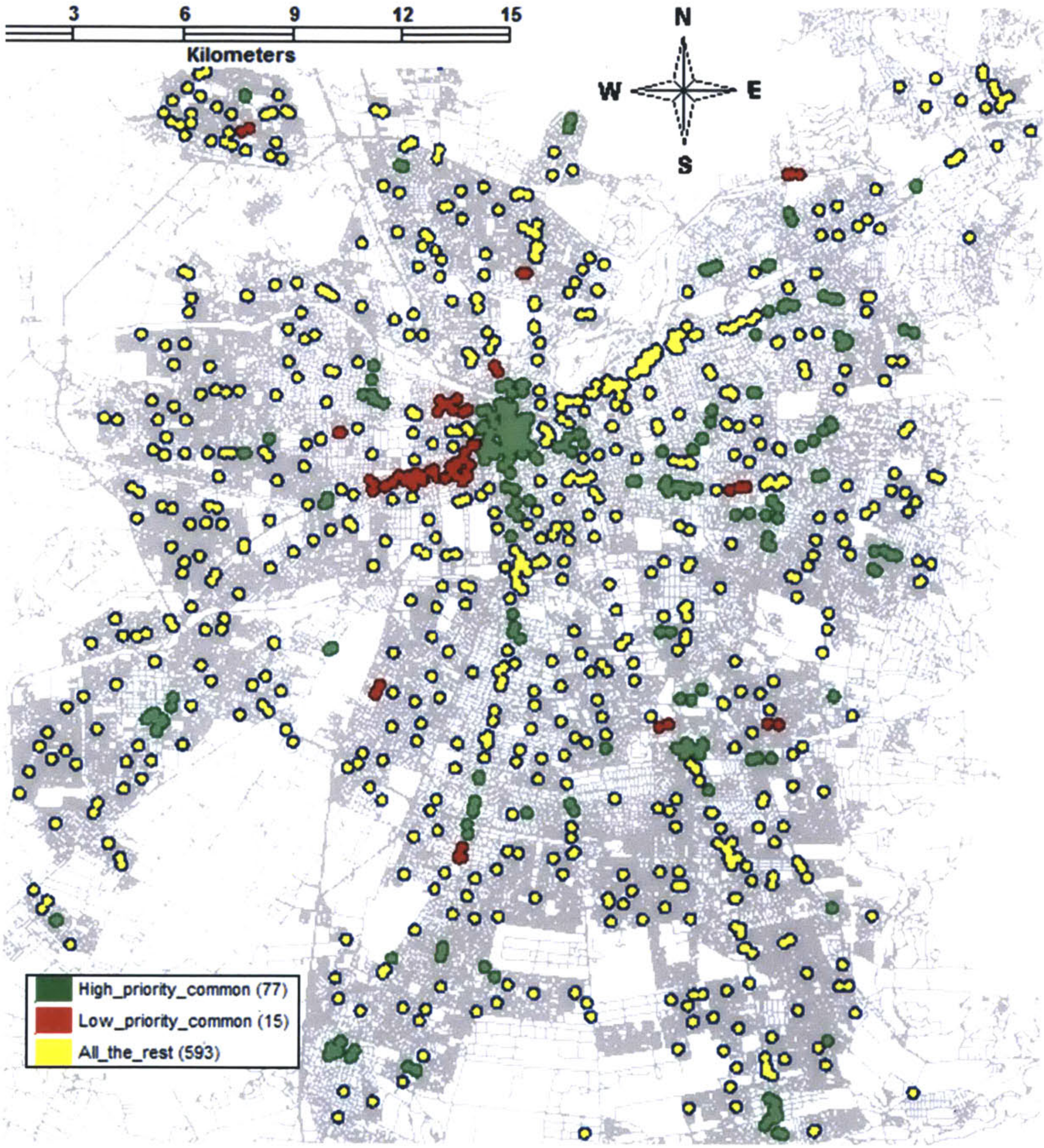

Figure 38: Commonality in Geoclusters

Finally, the methodology proposed in this thesis is useful not only to prioritize areas or geoclusters within cities; it also provides a traceable measurement of cities' level of saturation, allowing comparisons between territories, product categories and even different companies or industries.

Accordingly, in this case for example, a quick comparison between the cannibalization coefficients obtained for instant and lottery games delivers the following conclusions:

- It confirms the hypothesis of instant games' purchase process being more driven by impulse than lottery games.
- Lottery games' distribution network is nearer to its saturation point than instant games, although both currently have approximately the same number of stores. Given current demand conditions, therefore, lottery games' sales in this distribution channel seem to be closer to their potential than instant games.

The proposal, therefore, is to keep track of these coefficients, including posterior history incorporating new store openings and closings. One of the issues that hindered the significance of the coefficients in some situations was the lack of variability within geoclusters for the active number of stores. With additional history, particularly if store opening and closing campaigns are conducted based on the results of this analysis, the overall significance of the coefficients should increase. The described opening and closing campaigns would also make store gaps in each geocluster tend to zero, as stores would be opened in areas with opportunities, and closed in saturated ones.

From a research perspective, working with data from other industries in which the geographic extension of the distribution networks is bigger, offers interesting opportunities to test the reliability of the analysis. Accordingly, as mentioned before, one issue limiting the significance of the coefficients in some cases was the lack of enough variability in the number of active stores. The Lottery network includes approximately 1,200 stores in Santiago (the city where the analysis was done), so this was a somewhat foreseeable situation. Other industries, particularly ones intensive in the use of the traditional channel, have much more extensive networks. As examples, the ice-cream industry in the same territory covers approximately 8,000 stores and the soft drink industry exceeds 30,000. As the number of stores grows, so does the expected variability in the number of active stores, which could lead to results even more significant.

References

Achabal, D. D., Gorr, W. L., & Mahajan, V. (1982). MULTILOC: A multiple store location decision model. *Journal of Retailing,* 58(2), 5.

Craig, C. S., Ghosh, A., & McLafferty, S. (1984). Models of the retail location process: A review. *Journal of Retailing,* 60(1), 5.

De Neufville, R. (2011). *Flexibility in Engineering Design.* Cambridge: MIT Press.

Durvasula, S., Sharma, S., & Andrews, J. C. (1992). STORELOC: A retail store location model based on managerial judgments. *Journal of Retailing,* 68(4), 420.

Garee, M. L., & Schori, T. R. (1998). Modeling can help predict franchise "cannibalization." *Marketing News,* 32(24), 4-4.

Ghosh, A., & McLafferty, S. L. (1987). *Location strategies for retail and service firms*. Lexington, Mass.: Lexington Books.

Kaufmann, P. J., & Rangan, V. K. (1990). A model for managing system conflict during franchise expansion. *Journal of Retailing,* 66(2), 155.

Kimes, S. E., & Fitzsimmons, J. A. (1990). Selecting Profitable Hotel Sites at La Quinta Motor Inns. *Interfaces,* 20 (March-April), 12-20.